Spring 1-1-2012

Predictive Modeling of Metagenomes

Dan Brooke Knights
University of Colorado at Boulder, daniel.knights@colorado.edu

Predictive Modeling of Metagenomes

by

Dan Knights

B.S., Middlebury College, 2001

M.S., University of Colorado, 2010

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Computer Science

2012

This thesis entitled:

Predictive Modeling of Metagenomes

written by Dan Knights

has been approved for the Department of Computer Science

_____

Professor Michael C. Mozer

_____

Associate Professor Rob Knight

Date_____

The final copy of this thesis has been examined by the signatories, and we

Find that both the content and the form meet acceptable presentation standards

Of scholarly work in the above mentioned discipline.

Knights, Dan (Ph.D., Computer Science)

Predictive Modeling of Metagenomes

Thesis directed by Professor Michael C. Mozer

Human-associated microbial communities have been implicated in a variety of chronic diseases, including inflammatory bowel diseases, obesity, and autoimmune disorders like diabetes. Environmental communities are also important for bioconversion of waste products in biofuel production. However, microbiomes are highly complex systems involving mutualism and competition between many constituent organisms, and a variety of fundamental and interesting computational challenges remain in the modeling of pathogenicity and community-wide response to perturbations [1, 2]. In this thesis we discuss several computational and statistical approaches to predictive modeling of microbiome behavior using high-throughput metagenomic and transcriptomic sequencing data, including models that leverage biological structures such as phylogenies and gene ontologies to help extract features and constrain model complexity. We also demonstrate several applications of these approaches to real biological problems.

We successfully apply predictive modeling to new studies of human-associated and environmental microbial communities in several interdisciplinary collaborations with colleagues at numerous institutions around the world. These include a prominent study of the species and genes present in diverse mammalian gut communities, a study of the effects of yogurt consumption on gut microbial taxa and gene expression (i.e. transcriptomics) in humans and mice, and a large cross-sectional global survey of the human gut microbiota in

varied populations. We also develop SourceTracker, a Bayesian approach to predictive modeling of mixtures of microbial communities [3] with important applications in forensics, pollution studies, public health, and detection of sample contamination.

This dissertation introduces predictive modeling of human-associated and environmental microbial communities, increasing our ability to understanding the diversity and distribution of the human microbiota, and especially the systematic changes that occur in different physiological and disease states. We expect this type of predictive modeling to have far-reaching effects on health and disease [4].

Acknowledgements

I thank my dissertation co-advisors, Mike Mozer and Rob Knight, who have been highly influential in my development as a researcher. Their encouragement, ideas, and support are reflected throughout this thesis. Thanks to Aaron Clauset, Robin Dowell, and Diana Nemergut for their service on the thesis committee. Much of this work was done in collaboration with various members of the Knight lab and with collaborators at various labs around the country and around the world. Their names are included where appropriate in the text. This work was also supported by the United States National Institutes of Health (R01HG4872, R01HG4866, U01HL098957 and P01DK78669), the Crohn's and Colitis Foundation of America and the Howard Hughes Medical Institute. Thanks to my parents Ed Knights and Lynn Courtney for setting me on the path of science at an early age. Thanks to my sister Liz Knights for her advice and support. And, especially, many thanks to my wonderful wife Gina for her support and her unwavering confidence in me during my graduate studies.

CONTENTS

CHAPTER

TABLES

FIGURES

CHAPTER 1

1 Introduction[1]

Different people harbor radically different microbial communities, which likely play key roles in a wide range of chronic diseases. If we can identify groups of bacterial taxa present in a human body habitat that are consistently predictive of host phenotype for different illnesses or treatments, then these biological signatures can be used to build models that predict therapeutic outcomes based on an individual's specific microbiota. This approach, based on predictive models, has implications for diverse diseases that may benefit by modulation of the microbiota (e.g. through pre-biotics, pro-biotics, or targeted antibiotics), such as Inflammatory Bowel Diseases (IBD), obesity, diabetes, or diseases that are associated with malnutrition. Furthermore, given the recent finding that humans leave a signature of a distinctive skin microbiota on their keyboards [5], this work also has implications for forensic identification. The crux of the problem is coping with the complexity and high-dimensionality of human-associated microbiota. Some progress has been made towards establishing the feasibility of supervised classification of these communities [1], but there has been limited development of novel approaches, and many challenges remain. We discuss several of these challenges and important areas for future research into predictive modeling of human-associated microbial communities, as well as the potential applications that motivate this research.

---

[1] From: Knights D, Costello EK, Knight R. (2011). "Supervised Classification of Human Microbiota". FEMS Microbiology Reviews 35, 343-359.

[2] From: Knights D, Costello EK, Knight R. (2011). "Supervised Classification of Human

This dissertation introduces predictive modeling of human-associated and environmental microbial communities, transforming the state of the art in the field from a general demonstration that groups of communities differ from one another into a detailed understanding of how those groups differ. Understanding the diversity and distribution of the human microbiota, and especially the systematic changes that occur in different physiological and disease states, is predicted to have far-reaching effects on health and disease [4].

After an introduction to microbiome data, the dissertation begins by demonstrating the feasibility of the approach of predictive modeling of microbiomes. We first evaluate the performance of a suite of existing supervised classification algorithms using several real human-associated microbiome studies as benchmark classification tasks, and show that predictive modeling was successful on several of these tasks [1]. In this work we also demonstrated that a number of natural biological structures inherent in microbiome data can be leveraged to improve the accuracy of classifiers. A key finding was that the random forests classifier [6] consistently outperformed the other algorithms that we reviewed, likely due to its modeling of nonlinear effects and its ability to leverage many weakly predictive features.

Following these findings, we successfully apply predictive modeling to new studies of human-associated and environmental microbial communities in several interdisciplinary collaborations with colleagues at numerous institutions around the world. These include marker-gene surveys of wastewater treatment bioreactors [7], in which we discover that various sulfate-reducing members of the order Syntrophobacterales were the most discriminative of the different types of reactors. We are able to study the behavior of this reduced set of taxa in more detail, and we find that they were the most resilient to

environmental perturbations. We also contribute to a prominent study of the species and genes present in diverse mammalian gut communities by showing that the functional profile of a community (i.e. what genes are present) can be predicted directly from its phylogenetic profile using a nearest-neighbor approach [8]. In another study of the effects of yogurt consumption on gut microbial taxa and gene expression (i.e. transcriptomics) in humans and mice, we use supervised learning to demonstrate a generalizable difference between pre- and post-yogurt mice, and to identify highly discriminative genes [9]. These genes coded for enzymes involved in carbohydrate metabolism pathways that were also differentiated in the human gut, demonstrating that mouse models of the human gut may be effective in a translational medicine pipeline.

We also develop SourceTracker, a Bayesian approach to predictive modeling of mixtures of microbial communities [3]. SourceTracker uses a set of sampled training communities to characterize the distribution of taxa in suspected "source" environments. It then applies a stochastic sampling algorithm to estimate the proportion of a novel community contributed by each source environment, plus an unknown environment for any fraction of the community that is highly divergent from all of the sources. Previous work focused on detection of fecal contamination in water, mostly using predetermined indicator species and biomarkers [10-12]. SourceTracker's distinguishing features are its direct estimation of source proportions and its Bayesian modeling of uncertainty about the known and unknown environments. Community-wide source tracking has important applications in forensics, pollution studies, public health, and detection of sample contamination.

This work is driven by our interest in using statistical and computational models to help understand how and why human-associated and environmental microbial communities form highly varied and complex assemblages. Predictive modeling will be of paramount

importance in answering such questions, and we have therefore focused our research both on novel methodological development and on identifying important directions for future work in this area.

1.1 Introduction to microbiome data analysis

Recent advances in DNA sequencing technology have allowed the collection of high-dimensional data from human-associated microbial communities on an unprecedented scale. A major goal of these studies is the identification of important groups of microorganisms that vary according to physiological or disease states in the host, but the incidence of rare taxa and the large numbers of taxa observed make that goal difficult to obtain using traditional approaches. Fortunately, similar problems have been addressed by the machine learning community in other fields of study like microarray analysis and text classification. In this review we demonstrate that several existing supervised classifiers can be applied effectively to microbiota classification, both for selecting subsets of taxa that are highly discriminative of the type of community, and for building models that can accurately classify unlabeled data. To encourage the development of new approaches to supervised classification of microbiota, we discuss several structures inherent in microbial community data that may be available for exploitation in novel approaches, and we include as supplemental information several benchmark classification tasks for use by the community.

We now review some of the analyses that are typical of the current literature on the characterization of human microbiota, and then provide motivation for the application of supervised learning in this field of study. After introducing the benchmark classification tasks that we use in this review, we explore several possible constraints inherent in microbial community data that might aid researchers in choosing which type of models to employ. In many cases where appropriate models already exist, we demonstrate their

effectiveness by applying some examples to the benchmarks; in other cases we suggest directions for research into novel approaches.

Supervised classification requires training data, where each training sample has values for a number of independent variables, or features, and an associated classification label. In this review we demonstrate that the taxon relative abundance vectors from 16S rRNA gene sequence surveys can serve as useful input features for some classification problems. Many of the techniques that we discuss are also applicable to metagenomic surveys, where the input features would be the abundances of thousands of functional genes. Other measurements of microbial community configuration could serve as useful input features as well; many of these measurements have been described in detail in a prior review in this journal [13]. Typical results generated by human-associated microbial community analyses can be seen in [14-16], and often include the following components.

1.1.1 Analysis of taxon relative abundances

A common data structure in community ecological analysis is the sample-by-taxon abundance matrix. In addition to serving as the input for OTU-based measures of alpha and beta diversity (described below), these matrices can be mined for taxa whose relative abundances vary significantly with sample type or treatment. For example, one robust finding involving differences in taxon relative abundances has been the association of obesity with gut microbiota that have a lower relative abundance of bacteria from the phylum Bacteroidetes [17]. In this review, we use sample-by-taxon abundance matrices as training data in our benchmark classification tasks. A common feature of such matrices is their data sparseness: most taxa are confined to a relatively small fraction of samples (high endemicity). Other data types, discussed below, may also serve as useful inputs in future supervised classification tasks, but will not be analyzed directly in this review.

1.1.2 Alpha diversity analysis

Measures of alpha diversity (or, within-sample diversity) have a long history in ecology [18]. Alpha diversity scores have been shown to be differentiated for communities from several types of human body habitat. For instance, skin-surface bacterial communities have been found to be significantly more diverse in females than in males [19] and at dry sites rather than sebaceous sites [16, 20], and the gut microbiota of lean individuals have been found to be significantly more diverse than those of obese individuals [17]. These studies suggest that in some cases alpha diversity scores will be useful input features for building supervised classifiers.

1.1.3 Beta diversity analysis and clustering

Beta diversity analysis attempts to measure the degree to which membership or structure is shared between communities. Many classical metrics can be used to estimate the distance between communities, although those based on phylogenetic relatedness perform optimally in 16S rRNA-based surveys [13, 21]. A non-phylogenetic distance metric such as the common Euclidean distance treats all organisms as though they were equally related to one another, and thus it can fail to capture the similarity between two communities containing closely related organisms. This problem becomes especially important in microbial community analyses where individual species are not commonly shared across environments, such as on the human body.

Once measures of beta diversity have been calculated, the entire data set may be visualized using one of several ordination methods, such as non-metric multidimensional scaling or principal coordinates analysis (PCoA). PCoA performs a rotation of the inter-sample distance matrix (after centering) to represent those distances as accurately as possible in a small number of dimensions. Non-metric ordination has a similar goal, but

seeks to represent only the rank order of inter-sample distances, rather than the actual distances as in PCoA. After ordination, a reduced-rank approximation of the inter-sample distances can be visualized in two or three dimensions for exploratory analysis and for identifying samples that cluster by habitat or environmental factors. There is no reason to use only the *first* two or three dimensions, but the higher dimensions will represent increasingly subtle trends in the distance matrix. Another popular unsupervised method is to create a hierarchical clustering of samples, and to visualize the resulting tree. All of these approaches have the purpose of using a small number of dimensions to represent, as closely as possible, the actual differences between samples.

Numerous recent microbiota analyses have used sample clustering based on phylogenetic beta diversity metrics (e.g., UniFrac) to explore compositional similarities between communities. Correlations include, for example, diet and phylogeny in mammal guts [22], body habitat, individual, and time in healthy adults [20], and fingertip microbiota on touched surfaces [5]. The latter case is particularly notable because it suggests that supervised classification based on phylogenetic beta diversity might prove useful in future work in the field of forensic identification.

1.1.4 Introduction to Supervised Classification for Microbial Ecologists

Supervised classification is a machine learning approach for developing predictive models from training data. Each training data point consists of a set of input features (in this review, the relative abundance of taxa) and a qualitative dependent variable giving the correct *classification* of that data point. In microbiota analysis, such classifications might someday include disease states, therapeutic results, or forensic identification. The goal of supervised classification is to derive some function from the training data that can be used to assign the correct class or category labels to novel inputs (e.g., new samples), and to

learn which features (here, taxa) discriminate between classes. Common applications of supervised learning include text classification, microarray analysis, and other bioinformatics analyses. For example, when microbiologists use the Ribosomal Database Project (RDP) website to classify 16S rRNA gene sequences taxonomically, they are using a form of supervised classification (naïve Bayes) [23].

Machine learning methods are particularly useful for recognizing patterns in highly complex data sets such as human microbiota surveys. The human microbiota consists of about 100 trillion microbial cells, compared to our 10 trillion human cells, and these microbial symbionts contribute many traits to human biology that we would otherwise lack. For example, gastrointestinal microbes are involved in xenobiotic metabolism [24], dietary polysaccharide degradation [25, 26], immune system development [27], and a wide range of other functions. Compositional differences between microbial communities residing in various body sites are large, and comparable in size to the differences observed in microbial communities from disparate physical habitats [22]. Understanding the diversity and distribution of the human microbiota, and especially the systematic changes that occur in different physiological and disease states, is predicted to have far-reaching effects on health and disease [4].

Each sample in a typical study of microbiota (using second-generation sequencing technology) contains hundreds or thousands of DNA sequences from an underlying community consisting of thousands of unique species-level operational taxonomic units (OTUs, previous work includes a discussion of assignment of OTU clusters [28]). Ecological assessments of such surveys have generally been restricted to measuring taxon relative abundances, analyzing within- and between-sample diversity (alpha and beta diversity,

respectively), exploring beta diversity patterns using unsupervised learning techniques such as clustering and principle coordinates analysis (PCoA), and performing classical hypothesis testing. These approaches may be limited in their ability to classify unlabeled data or to extract salient features from highly complex and/or sparse data sets. Fortunately, many techniques in supervised learning are designed specifically for those purposes. For example, supervised learning has been used extensively with success in microarray analysis, a field with similar dimensionality issues, to identify small groups of genes that can be used to distinguish between different types of cancer cells [29]. These techniques may hold promise for future applications demanding a similar solution to microbial community classification, including medical diagnosis and forensics identification.

CHAPTER 2

2 Validation of supervised learning for microbiome analysis[2]

The main purpose of supervised learning is to build a model from a set of categorized data category labels can be any type of important metadata, such as the disease state of the host. The ability to classify unlabeled data is useful whenever alternative methods for obtaining data labels is difficult (as in the use of microbial communities from the human body in forensic identification [5]), or potentially fatal (as in the use of gene expression profiles to classify cancer types [29]). In this review we generally restrict our discussions to classification problems where the labels are discrete (qualitative), but much of the content is applicable to regression problems where the labels are continuous (quantitative).

This goal of building *predictive* models is very different from the traditional goal of fitting an explanatory model to one's data set; here we are concerned less with how well the model fits our particular set of training data, but rather with how well it will generalize to novel input data. Hence we have a problem of *model selection*: we don't want a model that is too simple or general, because it will fail to capture subtle but important information about our independent variables ("underfitting"), but we also don't want a model that is too complex or specific, because it will incorporate idiosyncrasies that are specific only to our particular training data ("overfitting"). What we really want to optimize is the *expected* prediction error (EPE) of the model on future data. An extensive introduction to model selection and supervised learning has been published previously [30].

---

[2] From: Knights D, Costello EK, Knight R. (2011). "Supervised Classification of Human Microbiota". FEMS Microbiology Reviews 35, 343-359.

When the labels for our data are easily obtained, as in the classification of microbiota by body habitat where the body habitat is known [20], we have no use for a predictive model. In these cases supervised learning can still be useful for building *descriptive* models of the data, especially in data sets where the number of independent variables or the complexity of their interactions diminishes the usefulness of classical univariate hypothesis testing. Examples of this type of model can be seen in the various applications of supervised classification to microarray data [29], in which the goal is to identify for further investigation a small but highly predictive subset of the thousands of genes profiled in an experiment. In microbial ecology, the analogous goal is to identify a subset of predictive *taxa*. Of course in these descriptive models accurate estimation of the EPE is still important; that is how we know that the association of the selected taxa with the class labels is not just lucky or spurious. This process of finding small but predictive subsets of features, called *feature selection*, will be of increasing importance as the size and dimensionality of microbial community analyses continues to grow.

A common way to estimate the EPE of a particular model is to fit the model to a subset (say, 90%) of our data and then test its predictive accuracy on the other 10% of our data. This gives us an idea of how well the model would perform on future data sets were we to fit it to our entire current data set. To improve our estimate of the EPE we can repeat this process ten times so that each data point is part of the held-out validation data once. This procedure, known as cross-validation, allows us to compare models that use very different inner machinery or different subsets of input features. Of course if we try many different models and select the one that gives us the lowest cross-validation error for our entire data set, it is likely that our reported EPE will be too optimistic. This is similar to the problem of making multiple comparisons in statistical inference; some models are bound to get “lucky”

on a particular data set. Hence whenever possible we want to hold out an entirely separate test set for estimating the EPE of our final model, *after performing model selection*. We do just that for the benchmarks used in this chapter: we randomly choose a fraction of the data to act as the test set; we use cross-validation *within* the remaining training set to perform model selection; and we report the prediction error of the final model when applied to the test set.

Even if we have established how to select the best parameters or degree of complexity for a particular kind of model, we are still faced with the problem of choosing what general class of models is most appropriate for a particular data set. The crux of choosing the right models for microbiota classification is to combine our knowledge of the most salient constraints (e.g., data sparseness) inherent in the data with our understanding of the strengths and weaknesses of various approaches to supervised classification. If we understand what structures are inherent in our data, we can then choose models that take advantage of those structures. For example, in the classification of microbiota, we may desire methods that can model non-linear effects and complex interactions between organisms. Or, due to the highly diverse nature of many microbial communities on the human body [20], we might want models designed specifically to perform aggressive feature selection when faced with high-dimensional data. Specialized *generative* models, discussed later in this review, can be designed to incorporate prior knowledge about the data as well as the level of certainty about that prior knowledge. Instead of learning to predict class labels based on input features, a *generative* model learns to predict the input features themselves. In other words, a generative model learns what the data "looks like", regardless of the class labels. One potential benefit of generative models such as topic models [31] and deep layered belief nets [32] is that they can extract useful information even when the data

are unlabeled. We expect the ability to use data from related experiments to help build classifiers for one's own labeled data to be important as the number of publicly available microbial community data sets continues to grow.

So far there has been almost no application of machine learning classification techniques to microbial community data, according to an extensive literature search. One exception is an analysis of soil and sediment samples [33], in which the authors classified the samples according to environment type using support vector machines (SVMs) and k-nearest neighbors (KNN). Their data generally classified well, with an expected prediction error (EPE) of 0.04 for a set of Idaho soil samples, and an EPE of 0.14 for Chesapeake Bay samples, although they characterized communities using amplicon-length heterogeneity profiles rather than 16S rRNA-based taxon abundances or alpha/beta diversity measures. In contrast, supervised learning has been used extensively in other classification domains with high-dimensional data, such as macroscopic ecology [34], microarray analysis (see above references) and text classification.

2.1 Benchmarks

While we do not intend this chapter to be a comprehensive review of classification techniques, we do want to demonstrate that supervised classifiers can be effective and useful in microbiota analyses. To this aim we used five benchmark classification tasks of varying size and difficulty involving actual human microbial communities. These data sets are included as supplemental information for the comparative evaluation of future approaches to supervised learning in this field. They are taken from two recent studies of human-associated microbial communities [5, 20]. Both data sets were 16S rRNA surveys sequencing the V2 region with 454 pyrosequencing. After denoising each data set with the PyroNoise algorithm [35], we used the default settings in the QIIME software package [36]

to pick OTU clusters with UCLUST [37] at a sequence similarity threshold of 97%. The choice of similarity threshold can have a significant effect on the quality of OTU abundances as predictive features, as we discuss later. In order to control for variable sequencing effort between samples we performed a single rarefaction at the depth of the shallowest sample. There are several other preprocessing steps that require parameterization; we used the default settings in QIIME, but a thorough benchmarking of the effects of various preprocessing choices on downstream analysis would be useful as a separate investigation.

Alpha- and beta-diversity analyses of the data are also likely to provide useful features for classification, although we constrain our discussion in this review to OTU abundances. For researchers wishing to perform a *de novo* analysis on these data sets, both have been made publicly available by the authors of the original studies. We now give details about the origin and purpose of each benchmark; Table 1 gives a summary of their sample sizes and dimensionality.

Benchmark	Training samples	Test samples	No. OTUs	No. classes
Costello et al. Body Habitats (CBH)	415	207	2741	6
Costello et al. Skin Sites (CSS)	268	133	2227	12
Costello et al. Subject (CS)	96	48	1592	7
Fierer et al. Subject (FS)	68	33	565	3
Fierer et al. Subject × Hand (FSH)	68	33	565	6

Table 1. Summary of benchmark data sets used in this chapter. Singleton OTUs were removed.

2.1.1 Benchmark 1: Costello et al. Body Habitats (CBH).

As noted above, microbial community composition tends to be highly differentiated between body habitats. The Costello et al. data included sample communities from 6 major categories of habitat: *External Auditory Canal (EAC), Gut, Hair, Nostril, Oral cavity,* and *Skin*. This benchmark is an example of a relatively easy classification task due to the generally pronounced differences between the communities, although some of the categories, such as *Hair*, are relatively underrepresented. The benchmark excludes samples from communities that were transplanted from another subject or body site. We are subject here to the choice by the original authors to separate *Hair* and *Nostril* from *Skin*, when the three categories seem to be easily confused by the classifiers that we review. Of course in practice these data would not normally require the use of a predictive model for classification, since the site of sampling would most likely be known. A more useful application for machine learning in this type of task is to perform feature selection to identify OTUs that are highly discriminative of the type of sampling site.

2.1.2 Benchmark 2: Costello et al. Skin sites (CSS).

This benchmark is a subset of the full Costello et al. data, containing only those non-transplant samples taken from skin sites. The class labels are the specific type of skin site, and contain 12 unique classes (e.g. *volar forearm, plantar foot, forehead, palm*, *etc.*). The compositional differences between these categories are generally much more subtle than in the CBH benchmark, so the classification task is more difficult. As with the CBH benchmark, predictive models aren't likely to be necessary for this particular classification task; the benchmark is instead intended to serve as a test bed for developing feature selection techniques as well as predictive techniques for use in other data sets where the category labels are more expensive to obtain.

2.1.3 Benchmark 3: Costello et al. Subject (CS).

This benchmark contains only a set of samples taken from the arms, hands, and fingers, excluding any "transplant" samples. The class labels are the (anonymized) identities of 7 of the 9 subjects in the study. We omitted two of the subjects, "M5" and "M6", because they had very few samples. This benchmark is moderately challenging due the fact that samples come from heterogeneous time points (84 from June of 2008, 28 from September of that year). Costello et al. observed significant variation in individuals over time, and indeed although several classifiers achieved perfect expected accuracy when trained and tested only within the June samples, the lowest error achieved in our mixed test set was 0.062. In this case, predictive models (as opposed to descriptive models) may be more directly meaningful than in benchmarks CBH and CSS above: the ability to classify individuals by their microbiota could have the same applications in forensics as in the Fierer et al. data set discussed next.

2.1.4 Benchmark 4: Fierer et al. Subject (FS).

This benchmark contains all samples from the Fierer et al. "keyboard" data set [5] for which at least 397 raw sequences were recovered (397 was chosen manually in order to include as many samples as possible). The class labels are the anonymized identities of the three experimental subjects, as with the CS benchmark above. This classification task is the easiest of all five benchmarks because of the clear distinctions between the individuals, because all of the samples come from approximately the same time point, and because of the large number of training samples available for each class.

2.1.5 Benchmark 5: Fierer et al. Subject × Hand (FSH).

This benchmark is a more challenging version of the previous one. The class labels are the concatenation of the experimental subject identities and the label of which hand

(*left* versus *right*) the sample came from on that individual. There were 3 subjects, so there are 6 classes in this benchmark.

Test sets: For each of the five benchmarks, we have created ten random splits of the data into training and test sets. A test sets contains 1/3 of the data for a given benchmark, and the proportion of each class in the test set is approximately the same as its proportion in the overall data set. The indices of the test sets that we used in this chapter are included with the benchmarks in the supplementary data.

2.2 Classifying classifiers

Many attempts have been made to review and organize published approaches to feature selection in high-dimensional classification problems, in some cases specifically with respect to microarray analysis, including, but not limited to [29, 38-41]. Lal et al provide an excellent paradigmatic framework for categorizing and discussing the available techniques [42]. The following section mentions a few issues in the design and application of classification methods relevant to microbial ecology; for a thorough taxonomy of classifiers, we refer the reader to the above articles.

2.2.1 Multiclass versus binary

An important feature of a classifier is whether it can easily support multi-category (multiclass) classification. Some models, such as the original support vector machine, inherently support only binary decision problems. Other methods, such as k-nearest neighbors, multinomial logistic regression and discriminant analysis allow for direct inference of multiclass decision boundaries. Binary classifiers can still be made to perform multiclass classification by collecting votes from one-versus-one (pairwise) or one-versus-all

classifiers, but the lack of multiclass support becomes problematic when the number of classes is high, or when the data set is large.

2.2.2 Approaches to feature selection

As discussed earlier, the goal of feature selection is to find the combination of the model parameters and the feature subset that provides the lowest expected error on novel input data. We consider feature selection to be of utmost importance in the realm of microbiota classification due to the generally large number of features (i.e., constituent species-level taxa): in addition to improving predictive accuracy, reducing the number of features we use will help us to produce more interpretable models. Approaches to feature selection are typically divided into three categories: filter methods, wrapper methods, and embedded methods.

As the simplest form of feature selection, filter methods are completely agnostic to the choice of learning algorithm being used; that is, they treat the classifier as a black box. Filter methods use a two-step process: Perform a univariate test (e.g. t-test) or multivariate test (e.g. a linear classifier built with each unique pair of features) to estimate the relevance of each feature, and select (a) all features whose scores exceed a predetermined threshold, or (b) the best n features for inclusion in the model; then run a classifier on the reduced feature set. The choice of n can be determined using a validation data set or cross-validation on the training set.

Although filter methods may seem inelegant from a theoretical viewpoint due to their inherent lack of optimality, they are used extensively in the literature. They have several benefits, including their low computational complexity, their ease of implementation, and their potential, in the case of multivariate filters, to identify

important interactions between features. The fact that the filter has no knowledge about the classifier is advantageous in that it provides modularity, but it can also be disadvantageous, as there is no guarantee that the filter and the classifier will have the same optimal feature subsets. For example, a linear filter (e.g. correlation-based) is unlikely to choose an optimal feature subset for a non-linear classifier such as a support vector machine or a random forest.

Wrapper methods are usually the most computationally intensive and perhaps the least elegant of the feature selection methods. A wrapper method, like a filter method, treats the classifier as a black box, but instead of using a simple univariate or multivariate test to determine which features are important, a wrapper uses the *classifier itself* to evaluate subsets of features. This leads to a computationally intensive search: an ideal wrapper would re-train the classifier for all feature subsets, and choose the one with the lowest validation error. Were this search tractable, wrappers would be superior to filters because they would be able to find the optimal combination of features and classifier parameters. The search is, however, not tractable for high-dimensional data sets, so the wrapper must use heuristics during the search to find the optimal feature subset. The use of a heuristic limits the wrapper's ability to interact with the classifier for two reasons: the inherent lack of optimality of the search heuristic, and the compounded lack of optimality in cases where the wrapper's optimal feature set differs from that of the classifier. We do not consider wrappers further in this review, since in many cases the main benefit of using wrappers instead of filters, namely that the wrapper can interact with the underlying classifier, is shared by embedded methods (discussed next), and the additional computational cost incurred by wrappers therefore makes such methods unattractive.

Research on embedded feature selection techniques has been plentiful in recent years, for example in [42-44]. Embedded approaches to feature selection perform an integrated search over the joint space of model parameters and feature subsets so that feature selection becomes an integral part of the learning process. Embedded feature selection has the advantage over filters that it has the opportunity to search for the globally optimal parameter-feature combination. This is because feature selection can be done with knowledge of the parameter selection process, whereas filter and wrapper methods treat the classifier as a "black box". As discussed above, performing the search over the whole joint parameter-feature space is generally intractable, but embedded methods can use knowledge of the classifier structure to inform the search process, while in the other methods the classifier must be built from scratch for every feature set.

The classifiers discussed in this chapter include several that perform embedded feature selection, several that employ filter methods, and several that perform no explicit feature selection at all but are nonetheless effective in high-dimensional data. The rest of this review is organized by several characteristics of microbial communities that we believe are important to consider when choosing between classification techniques. In cases where applicable techniques exist we review several published examples; in other cases we suggest directions for future work.

2.3 Feature extraction

Sometimes referred to as feature induction, *feature extraction* is the process of creating or learning useful transformations of the original set of input features. Certain rigid rule-based approaches such as decision trees or rule induction can derive features that are directly interpretable to the end user. In other cases the derived features can model complex interactions between the observed input features. In the case of a layered model

such as a deep belief net [32], each level of derived features tends to model more abstract concepts than the previous. For example, in the process of learning to recognize objects in images, the first layer of a deep belief net might produce derived features that correspond to detected edges, the next to detected visual features, and the deepest to actual detected objects. For the purposes of classifying microbial communities, feature extraction is beneficial if it can (a) increase the classification accuracy, or (b) provide useful insight into the structure of our data.

It may not always make sense to do feature selection (section 4.2) with OTU counts, because in some cases OTUs may be relatively exchangeable with one another in terms of functional behavior within the same type of community. A feature extraction method such as a deep belief net could have the opportunity to learn either/or relationships, whereas a feature selection technique might be forced to choose between partially exchangeable organisms for inclusion in a classifier.

2.4 Review of selected classifiers

Microbial community data tend to be sparse; Figure 1 shows a histogram of the frequency at which OTUs were observed in a given portion of samples in the Costello et al data set. The full dataset consists of 816 samples and yields 14,254 OTUs when sequences are clustered with UCLUST at 97% similarity. The histogram in the figure excludes singletons, of which there were 10,471. Only 131 (0.9%) of the 14,254 observed OTUs were present in more than 10% of the samples; 97.7% of the species abundance matrix entries were zeros. Such high levels of sparsity are common in 16S rRNA microbial surveys, although the number of unique OTUs observed depends on several factors. The original sequencing data consists of millions of (generally) unique DNA sequences. In common practice, these sequences are binned into clusters at a pre-determined similarity threshold.

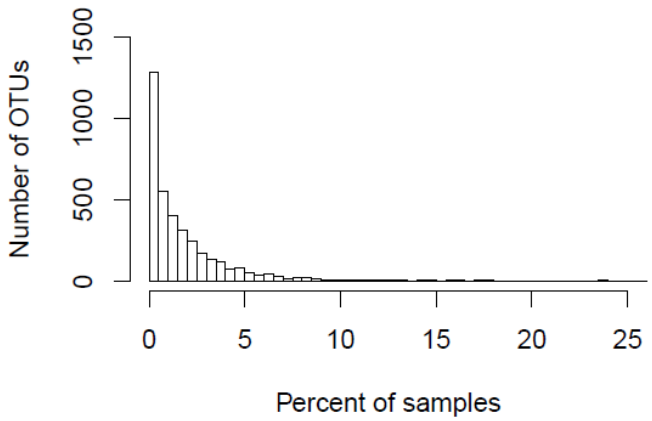

Figure 1. Histogram of number of OTUs present in a given percentage of samples.

Frequencies are for the full Costello et al. data set. Data were denoised using the PyroNoise algorithm, and OTUs were then picked at 97% similarity with the UCLUST software package. Of the 14,254 OTUs, 10,471 singletons were excluded before producing this histogram. This exemplifies the extreme sparsity typical of microbial community data sets.

However, as shown in [35], second-generation sequencing methods can be quite error-prone, and aggressive de-noising is required to avoid having a falsely high number of so-called "distinct" taxonomic groups, so the OTUs in the Costello data shown in Figure 1 were obtained after de-noising with the PyroNoise algorithm. The choice of bin size (similarity threshold) for the step of picking OTUs has a notable effect on the discriminative value of the resultant OTUs, as will be discussed later in this chapter.

The inherent sparseness of the abundance matrix is the fundamental challenge in building classifiers for microbiota; as discussed earlier we know that systematic differences exist between many types of communities (such as between the gut communities of different human subjects), but identifying which OTUs will produce both good discrimination within the training data and good generalization to future test data remains challenging when so few OTUs are actually shared across communities. Although the following models have no explicit mechanism for incorporation of other kinds of prior knowledge, they are expected to perform well in high-dimensional classification problems

such as those of concern, based on their published performance in other tasks with similar sparsity and dimensionality.

2.4.1 Random forests

Although the random forest (RF) classifier is not explicitly designed for performing feature selection or dimensionality reduction, it is one of the top performers in microarray analysis [29] as well as in many other domains with high-dimensional data [6]. Random forests are an extension of *bagging*, or bootstrap aggregating, in which the final predictions of the model are based on an ensemble of weak predictors trained on bootstrapped samples of the data. A random forest consists of many such classifiers, each of which is a decision tree. At each level of the decision tree several randomly weighted linear combinations of small randomly selected subsets of features are evaluated by their ability to discriminate between categories, and the best subset is chosen to perform the split at that node. Other methods may outperform random forests in the presence of large numbers of irrelevant features [45], but the strong performance of random forests in microarray analysis indicates that they should be effective for classifying at least moderately sized microbial communities. One drawback to RF is that it does not explicitly perform feature selection. It does provide a natural ranking of the relative importance of features [6], but since most features are given at least some non-zero importance score we cannot easily identify the smallest number of features required to maintain a given level of accuracy.

2.4.2 Nearest shrunken centroids

The nearest shrunken centroids classifier (NSC) [43] performed well on microarray data in an extensive comparative review [29]. It is also fast, with algorithmic complexity scaling linearly in the number of features. NSC begins with the simplifying assumptions that an OTU's relative abundance is approximately normally distributed within each class,

and that its abundance is independent of the abundance of other OTUs. Of course in general we might want to model the covariance of OTUs, but it may do more harm than good when we have many more OTUs than data points; in such cases there are likely to be spurious correlations due to chance. In this simple model we would simply find, for each OTU, the mean within each class and the pooled within-class variance. This would give us an estimate of the location and spread of the centroid of all OTUs in each class. To classify a new sample we would then calculate the log likelihood in each class of that sample's OTU abundance vector given the class centroids and the normality assumption, and then choose the class with the highest log likelihood.

Without any modification, this model is simply a linear discriminant analysis that assumes no covariation between OTUs (i.e. diagonal covariance). To instead perform feature selection and effectively denoise the centroid for each class, we first find the z-scores of the OTUs in each class centroid relative to the overall centroid. We then shrink all of these z-scores by a fixed amount λ, causing any with an absolute value of less than λ to become zero (i.e. we apply soft thresholding). The value of λ can be chosen using validation data or cross-validation within the training data. This gives us new shrunken z-scores for each OTU in each class. We map these back onto the overall centroid to get a shrunken centroid for each class, and then use these in place of the full centroids to classify new points as described above. The soft thresholding of the z-scores has the effect of zeroing out the least distinctive OTUs in each class. Hastie et al. note that this procedure is basically applying a lasso-style penalty (see elastic net, next) to the class z-scores [30].

2.4.3 The elastic net

The elastic net (ENET) is a powerful, theoretically well-founded classifier that performs embedded feature selection with support for regression and binary and multiclass

classification [44]. In an ordinary least squares regression, all of the regression coefficients are completely unconstrained and may take on arbitrarily large values. This can lead to highly variable and unreliable models when some of the features are correlated with each other, and can cause overfitting when the number of input variables greatly exceeds the number of data points (as in typical microbiota experiments). One way to combat this problem is to reduce the variance of the model by constraining the size of the regression coefficients. This approach is known as *regularization*, and the choice of constraint placed on the coefficients is known as a *penalty*. The ENET penalty is a hybrid between two common penalties, the "ridge" penalty, which constrains the L2-norm (sum of squared values) of the coefficients, and the "lasso" penalty, which constrains the L1-norm (sum of absolute values) of the coefficients. This allows the ENET to leverage the tendency toward sparseness (i.e. setting many coefficients to zero and thus performing feature selection) of the lasso penalty while retaining the capability of the ridge penalty to retain groups of correlated variables. Also, the inclusion of the L2 penalty term allows the model to retain, if necessary, more input variables than there are data points, a limitation when the L1 lasso penalty is used alone. Given a standard regression problem with standardized predictors and response variable, the elastic net loss function is defined as:

$$L(\alpha,\beta) = |y - \mathbf{X}\beta|^2 + \alpha |\beta|_1 + (1-\alpha)|\beta|^2$$

Where $|\beta|_1$ and $|\beta|^2$ are the L1 and L2 norms of the vector of regression coefficients, and $|y - \mathbf{X}\beta|^2$ is the sum of the squared residuals from the fit. This penalty model is called the "elastic net" because, according to the authors, it is like an elastic fishing net that stretches just enough to catch "all the big fish". The ENET penalty allows the model to find the optimal compromise between the L1 and L2 penalties, and the value of the mixing

parameter α can be chosen by performing cross-validation on the training data. For multiclass classification problems we perform multinomial logistic regression instead of linear regression.

The ENET multinomial classifier has been shown to perform well on microarray data [44]. The fact that the ENET is capable of retaining groups of correlated input variables augurs well for its application to the classification of microbial communities, because in general we expect that some organisms have correlated patterns of abundance across communities. In the Costello et al. benchmark data set, for example, each OTU is on average highly correlated or anti-correlated (Pearson's coefficient of greater than 0.5 or less -0.5) with 21.9 other OTUs (0.6%), and with 17.6 other OTUs (1.3%) in the Fierer et al. benchmark data set.

2.4.4 Support vector machines

Support vector machines (SVMs) also tend to be excellent all-around classifiers. The basic model was described previously [46]. Traditional SVMs are restricted to binary classification tasks, although they are commonly applied to multiclass tasks by breaking the task into separate binary one-versus-one or one-versus-all tasks, and then allowing each model to vote for the final classification. SVMs have been effective in microarray classification tasks [47]. The general approach taken by support vector machines is to embed the n data points in an $n\text{-}1$ dimensional space in which the classes are linearly separable, and then to identify the hyperplane (known as the maximum-margin hyperplane), that maximizes the gap between the classes. This has the effect of minimizing the generalization error on unseen data. Choosing the right spatial embedding can allow an otherwise nonlinear class boundary to become linear, but in the case where the data are still not linearly separable, the SVM finds the maximum *soft* margin, where the objective

function is penalized by some chosen cost function based on the distance of misclassified samples from the decision boundary. A support vector machine is so called because the separating hyperplane is supported (defined) by the vectors (data points) nearest the margin.

Although SVMs can perform poorly when given large numbers of irrelevant features, several approaches to feature selection combined with SVMs have proven successful in other high-dimensional classification problems, and hence these approaches may be useful ways to apply SVMs to microbial community data. In some studies filter methods such as the ratio of the between-class sum-of-squares to the within-class sum-of-squares (BSS/WSS) have been effective for classifying microarray data or text when combined with SVMs [29, 48]. Other approaches use embedded feature selection, such as the zero-norm SVM, or R^2W^2 feature selection [49]. More validation is needed for the zero-norm SVM, but it has been shown to perform well on one yeast classification experiment [50]. R^2W^2 is simple, performs well on microarray test data, and tends to use a small number of features relative to other feature selection methods, although it does not support native multi-category classification. In this review we use traditional SVMs both without filtering and with the three filter methods discussed next.

2.4.5 Filter methods

The purpose of a filter is to identify features that are generally predictive of the response variable, or to remove features that are noisy or uninformative. Forman evaluates many common filters including the between-class χ^2 test, information gain (decrease in entropy when the feature is removed), various standard classification performance measures such as precision, recall, and the F-measure, and the accuracy of a univariate

classifier, among others [39]. He also proposes a novel filter, called the Bi-Normal Separation (BNS), which treats the univariate true positive rate and false positive rate (*tpr*, *fpr*, based on document presence/absence in text classification) as though they were cumulative probabilities from the standard normal cumulative distribution function, and he uses the difference between their respective z-scores, $F^{-1}(tpr) - F^{-1}(fpr)$, as a measure of that variable's relevance to the classification task. This approach is noteworthy because it outperformed all other filter methods in almost every performance measure that Forman reviewed, across several hundred test data sets. According to him, the BNS is effective because of the type of decision boundary it creates in the positive/negative document space of low-frequency words. More specifically, when compared to other methods, it tends to be just aggressive enough in avoiding rare words with mildly predictive *tpr*-to-*fpr* ratios. The fact that a single filtering method outperformed the others on an extensive and diverse suite of hundreds of experiments implies that a similar review in the domain of microbiota analysis may be illuminating.

The BNS filter will likely require some adaptation before it can be used on microbial community data, due to its reliance on the presence/absence of the feature in a given sample. For example, it is known in some cases that frequently-occurring (i.e. non-sparse) microorganisms are associated with different clinical conditions [17]. Thus the fact that the BNS uses presence/absence for determining true and false positive rates could cause common but predictive OTUs to be ignored. It may be applicable in Forman's original formulation for extremely sparse datasets (high beta- and alpha-diversity). The approach we followed in this review was to build univariate multiclass classifiers for all features, and to find the average true positive and false positive rates for each feature. These rates were then used to score features using BNS.

The second filter we consider is a type of backward feature elimination called recursive feature elimination, which was tailored to the SVM (SVM-RFE) [51]. In SVM-RFE we train a classifier using the full set of features (OTUs), remove the feature with the least influence on the current margin, and repeat with the reduced feature set until all features have been removed. The features are then ranked by importance in reverse order of removal.

The third and last filter that we discuss is the simple BSS/WSS filter. BSS/WSS is common in the literature and has been demonstrated to be effective on non-linguistic domains such as microarray classification [29]. The BSS/WSS score of a feature *j* is defined as its ratio of between-group sum-of-squares to the within-group sum-of-squares:

$$\frac{BSS(j)}{WSS(j)} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{K} I(y_i = k)(\bar{x}_{kj} - \bar{x}_{\cdot j})^2}{\sum_{i=1}^{n}\sum_{k=1}^{K} I(y_i = k)(x_{kj} - \bar{x}_{kj})^2},$$

where K is the number of classes, n is the number of training samples, and $\bar{x}_{\cdot j}$ is the average value of the feature across all classes. Previous experiments demonstrate that the BSS/WSS generally performs well [29], and the results that obtained using the SVM with a radial basis kernel are comparable to those observed on the same data sets using embedded (non-filtered) SVM approaches [49].

2.5 Performance of selected classifiers on human microbiota

Table 2 contains the results of the unfiltered RF, NSC, ENET and SVM classifiers on all of the benchmark data sets. For the four classifiers we used publicly available implementations in the statistical software package *R*. Also included is the multinomial naïve Bayes (MNB) classifier, which is discussed later in the context of generative models.

Method	Mean rank	Mean increase in error	Average test error (Average number of OTUs)				
			Costello Habitats	Costello Skin Sites	Costello Subject	Fierer Subject	Fierer Subject × Hand
RF	1.7	.01	.09 (2484)	**.34** (2152)	**.11** (1522)	**.00** (475)	.28 (507)
MNB	2.3	.05	**.08** (2741)	.42 (2227)	.23 (1592)	.04 (554)	**.23** (554)
NSC	2.4	.04	.09 (1842)	.42 (2006)	.20 (1391)	.01 (320)	.25 (326)
ENET	3.6	.06	.11 **(385)**	.43 **(700)**	.13 **(566)**	.05 **(59)**	.33 **(137)**
SVM	5.0	.25	.19 (2741)	.55 (2227)	.54 (1592)	.17 (554)	.54 (554)

Table 2. Performance of various classifiers on the benchmark data sets.

For each classifier, for each benchmark, we show the mean test error over 10 repeated train/test iterations (standard errors not shown), and the average number of features used in the final models produced over the 10 train/test iterations. Each train/test iteration consists of training the model on a randomly selected training set (training set sizes shown in Table 1), and then recording that model's error in predicting the labels for the unseen test set. The "mean rank" column gives the average rank of that classifier across all benchmarks (lower is better); the rank of a classifier on a single benchmark is the standard fractional ranking. Fractional ranks were determined by considering models as tied when the better model's performance was within one standard error of the worse model's performance. The "mean increase in error" column gives the average difference between that model's test set error and the best model's test set error for a given benchmark (lower is better). Results in bold are within one standard error of the best result for that column.

For each benchmark we report the number of features used by the models; in the case of RF we show the number of features with a non-zero importance score.

Using the *randomForest* package [52] with default settings, RF achieves the best performance of all classifiers, with the highest rank (inclusive of ties) for every benchmark. To evaluate the NSC classifier we used the *pamr* package [53] with default settings. We see in Table 2 that NSC has fair performance on most of the benchmarks, but is clearly outperformed by the RF classifier in terms of test error and by the ENET classifier in terms of dimensionality reduction (i.e., reducing the number of OTUs required by the model). Using the ENET package *glmnet [54]*, and searching over 10 possible values for α

(0.01,.1,.2,…,0.9,1.0) with otherwise default settings, we found that the ENET had somewhat higher prediction error on average than RF. In most cases, however, it drastically reduced the number of features used for the classification, and we found that the OTU subsets selected by the ENET tended to be good features for the RF classifier. For example, the 367 and 27 OTUs selected by the ENET for the CBH and FS benchmarks, respectively, allowed the RF classifier to obtain at least as high accuracy as with the full set of OTUs. While we do not know in general if these classifiers tend to agree about which features are important, the RF, NSC, and ENET classifiers had reasonable overlap for the FS benchmark. Figure 2 shows a Venn diagram of the feature selection agreement between these three classifiers and the SVM-RFE filter (discussed below).

Figure 3 shows a heatmap plot of the 27 OTUs selected by the elastic net for the FS benchmark. Using these OTUs the RF classifier had 99.4% test accuracy across all test sets.

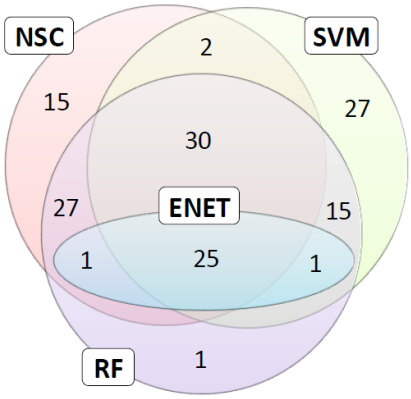

Figure 2. Venn diagram of OTUs selected by various filter methods.

Selected OTU subsets for the RF, NSC, and ENET classifiers and the SVM-RFE filter on the FS benchmark. For the RF and NSC classifiers and the SVM-RFE filter, we included only the top 100 features. For RF, these were chosen by the default RF "importance" score (Breiman 2001); for NSC, they were chosen by the average rank of the OTU during cross-validation as reported by the pamr package.

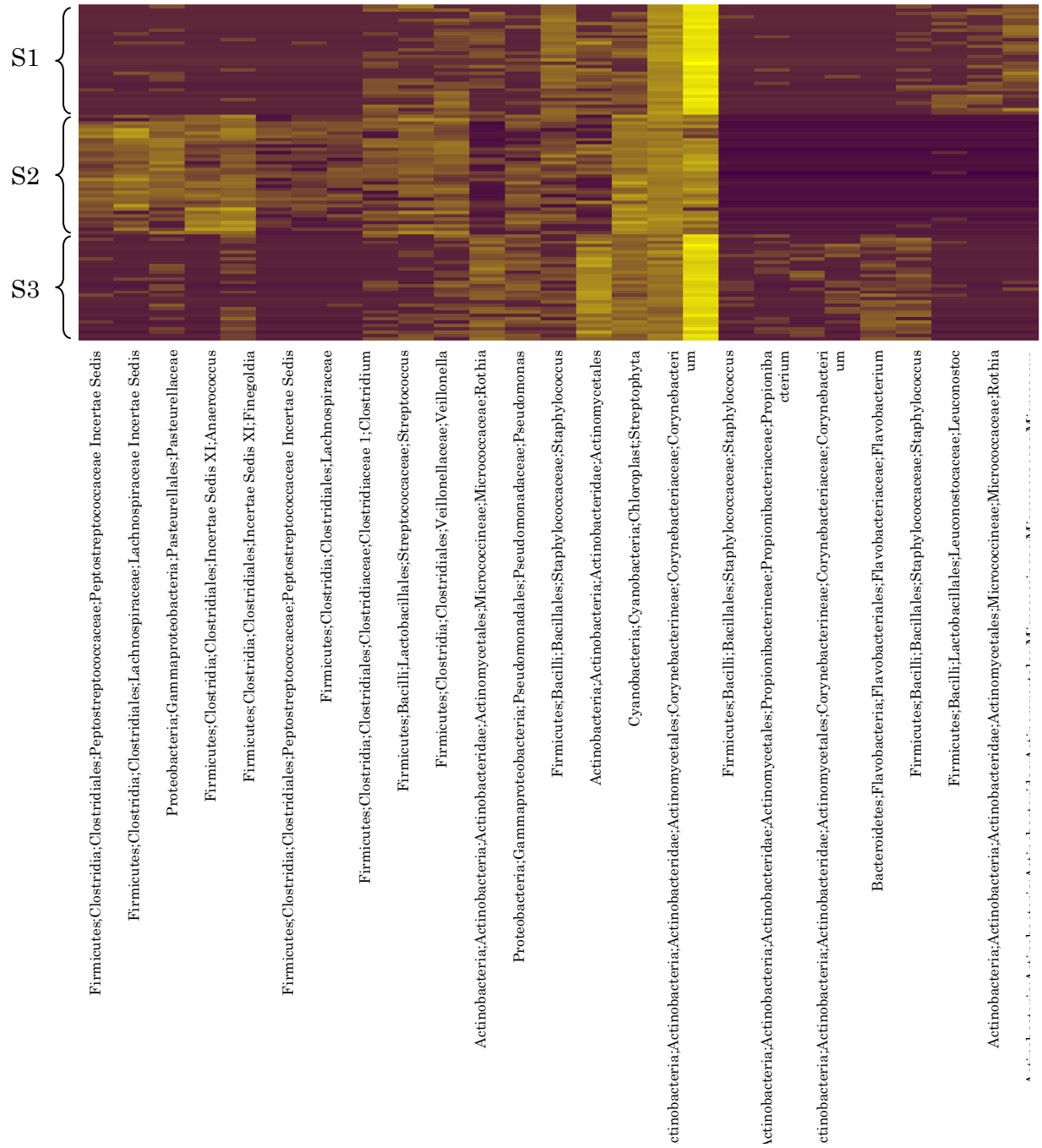

Figure 3. Heatmap of predictive OTUs for FSH benchmark.

Heatmap of the log relative abundance of 27 OTUs forming the unique microbial fingerprint of each subject in the FSH benchmark. The OTUs were selected by the elastic net classifier for assigning hand, fingertip, and keyboard microbial communities to the correct host. The elastic net parameters (α, λ) were tuned using 10-fold cross-validation; using those parameters, the final model was then trained on the entire data set. OTU lineages were assigned by the RDP classifier. Rows in the heatmap are standardized to zero mean and unit variance. Hierarchical clustering of columns was performed with Ward's method; rows were sorted by subject.

Thus these OTUs can be interpreted as representing the unique microbial "fingerprint" of each individual. In the heatmap we see interesting systematic difference between individuals. The OTUs chosen by the ENET are quite diverse; it seems that each individual has a unique representation of OTUs across many bacterial families. Some of these may be related to distinct types of non-keyboard surfaces that are commonly touched by each subject. For example, one subject appears to have a consistent over-representation of *Pasteurellaceae*, commonly found on mucosal surfaces of humans and animals [55]. Another has very high relative abundances of *Streptophyta*, a plant phylum. It is important to note that this subset of features is not likely to be optimal in size or choice of OTUs for minimizing EPE; finding such a subset is intractable for all but very small data sets. What we can say is that this is a highly predictive subset, capable of achieving perfect or near perfect accuracy on our benchmark test set.

For the CBH benchmark, the OTUs selected by the ENET are representative of the previous findings related to human body microbiota reviewed above. Notably, the *Oral cavity* samples are distinguished by their relative abundance of *Streptococcus*, *Pasteurellaceae*, *Prevotella*, and *Neisseria*, and as expected, *Bacteroides*, *Faecalibacterium*, and *Lachnospiraceae* tend to be over-represented in samples from the gut. This result is a validation of the utility of supervised classifiers for selecting relevant features in a descriptive model.

For reviewing SVMs we used the implementation in the "e1071" package in *R* [56] with default settings (and the radial basis kernel). To optimize the *cost* and *gamma* parameters of the SVM we performed a grid search over five values for each parameter and chose the combination that minimized cross-validation error within the given training set.

We found that SVMs had consistently poor performance on the benchmarks when used without filtering. However, when combined with the SVM-RFE filter, SVM achieved similar performance to RF, with dramatically smaller OTU subsets. The full results of the BSS/WSS, modified BNS, and SVM-RFE filters when combined with the RF and SVM classifiers are shown in Table 3. To obtain these filtered results we first ranked the OTUs by each filter method, and then used the top n OTUs to build our final classifier, where we selected the n that minimized cross-validation error within the training set. This approach has led to very small feature sets with excellent accuracy on similar data sets [51]. The SVM-RFE and modified BNS results were all within one standard error for both classifiers. Both classifiers performed better with a filter (Table 3) than without (Table 2), and a comprehensive study of filter methods applied to microbiota classification is recommended.

Filter	Classifier	No. features	Test error
SVM-RFE	SVM	40	.27
SVM-RFE	Random forests	34	.25
BSS/WSS	SVM	64	.41
BSS/WSS	Random forests	52	.28
BNS	SVM	70	.29
BNS	Random forests	70	.26

Table 3. Performance of classifiers with filters on the FSH benchmark.

For each classifier-filter combination including the BNS, BSS/WSS, and SVM-RFE filters, the number of features was selected by leave-one-out cross-validation on the training set; this table reports the number of features and the test set error.

2.6 Mining phylogenetic relationships

It is possible to use a global alignment of the DNA sequences belonging to the different OTUs in a collection of microbial communities to place those OTUs in a phylogenetic tree. This tree has the potential to provide much more information about the similarity of communities than the raw counts of OTUs, as the tree allows us to measure the similarity of two communities by how closely related their constituent taxa are. In contrast, using only the raw abundance of OTUs to calculate inter-community distance assumes that all OTUs are equally related to one another (i.e. related by a "star" phylogeny).

Phylogenetic distance measures that use the structure of the tree have been shown to recover known clusters of microbial communities in data sets where non-phylogenetic distance measures fail [57]. For example, Figure 4 shows sample scores on the first two PCoA axes of the inter-sample distances in the CBH benchmark using phylogenetic

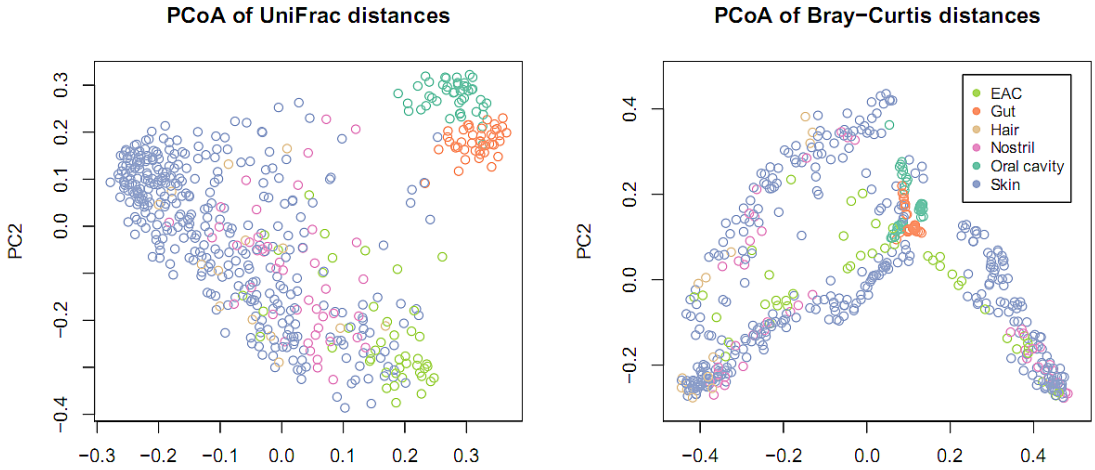

Figure 4. Comparison of phylogenetic and non-phylogenetic distance measures.

Left: the first two principal axes of principal coordinates analysis of body habitat samples from the CBH benchmark based on UniFrac (phylogenetic) distances; Right: the same analysis using Bray-Curtis (non-phylogenetic) distances.

(UniFrac) and non-phylogenetic (Bray-Curtis) distance metrics. The points in each plot represent individual microbial communities, and the colors represent the body sites from which the samples were taken. The phylogenetic distance metric clearly shows much better clustering of the samples by body site than the non-phylogenetic distance metric. Phylogenetic analysis is almost certain to provide useful derived features for supervised learning in some cases, although how best to mine the phylogenetic relationships for useful features is an open question.

2.7 Phylogenetic depth of OTUs

As discussed earlier, the raw data produced in a 16S rRNA-based survey consists of millions of (generally) unique nucleotide sequences. In order to facilitate analysis these sequences are commonly binned into clusters based on similarity at a pre-determined similarity threshold. In this chapter we use the default settings of the QIIME software package for picking OTUs [36]. By default QIIME employs UCLUST for picking OTUs at 97% sequence similarity, but the choice of similarity threshold may provide a natural source of dimensionality reduction: as we lower the similarity threshold, the bins get larger, and we get fewer OTUs. Figure 5 shows the average test error for the RF classifier on the FSH benchmark for each of 10 random train/test splits as we varied the level of similarity within OTU clusters. Quite surprisingly, the expected performance of the classifier is about the same at all levels of similarity between 65% and 95%, with almost a 100-fold range in dimensionality. That is, for the FSH benchmark a model built using 14 very general OTUs is just as effective on average (although with a bit higher variance across training sets) as a model built using 1,282 very specific OTUs. Also interesting is that for this classification problem, accuracy gets noticeably worse at very high levels of similarity such as 97% and

99%, suggesting that for some data sets, too much specificity makes it difficult to capture broad trends at higher taxonomic levels.

Clearly the issue of feature selection or dimensionality reduction in microbiota analyses is in some cases intimately tied to the taxonomic specificity of our OTUs. However, there are certain known phyla for which subtle genetic differences even between different strains of a species can make the difference between pathogen and non-pathogen, such as in the case of *Pseudomonas aeruginosa*, so it cannot be the case that we always want to reduce dimensionality by reducing taxonomic specificity. It may be that hybrid models using several levels of phylogenetic binning will outperform those constrained to any one bin size, and this is certainly an area that requires further research.

2.8 Metabolic functions as latent factors

OTUs that are relatively exchangeable with one another in terms of functional (metabolic) behavior may not be present in the same communities due to competitive exclusion [58]. Therefore it may not always make sense to do feature selection with OTUs, especially when we are dealing with highly specific OTU clusters. If indeed our classification categories are differentiated by the functional behavior of their communities rather than by the specific species-level taxa they comprise, then what we really want to learn is a set of functional equivalence classes, each containing a set of functionally redundant OTUs. We can then do inference in the reduced space of the latent functional profiles that are generating the observed community structures, rather than in the much more complex space of the OTUs

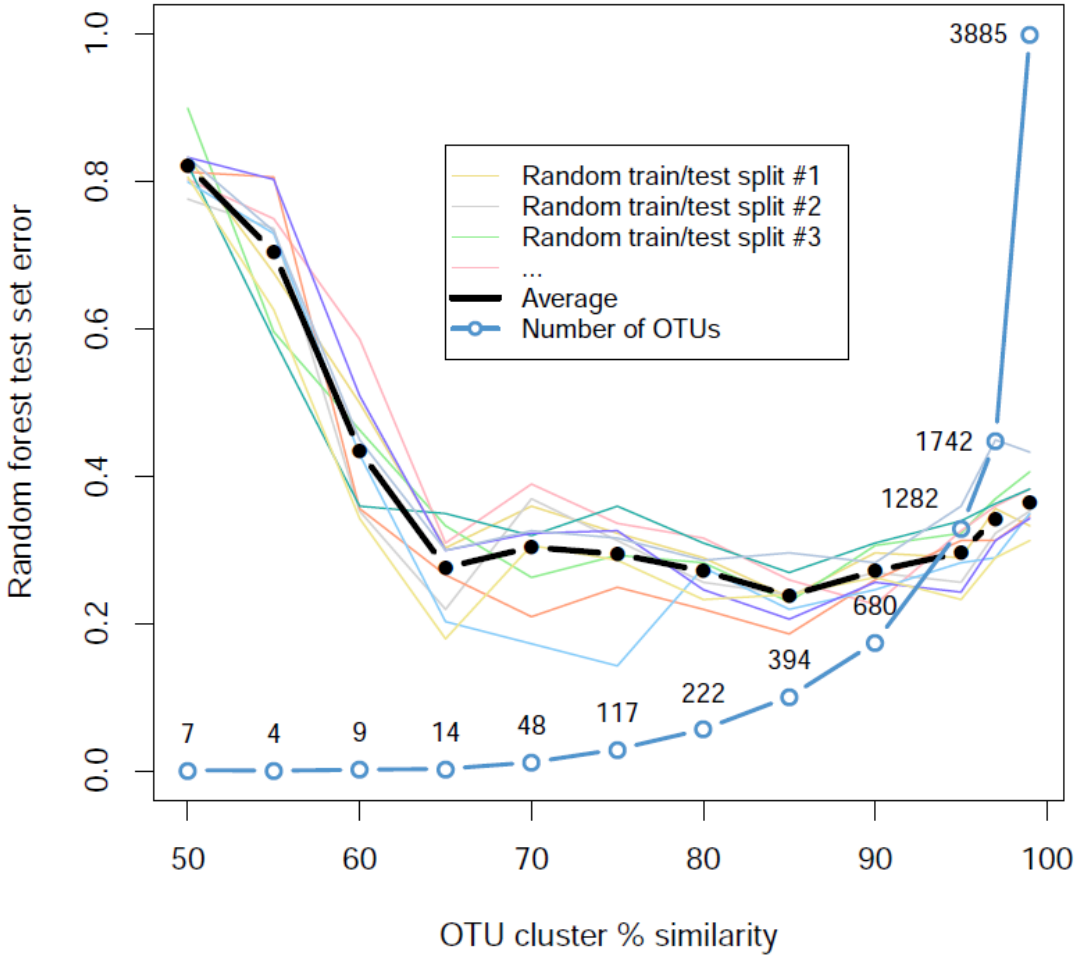

Figure 5. Prediction error versus OTU cluster specificity.

Random forests test set error as the percent similarity threshold for building OTU clusters is varied using the UCLUST software package. Colored lines show the results for ten randomly chosen splits of the data into training and test sets; the thick black line shows the average of all ten. Also shown (in blue triangles) is the number of OTUs chosen at each similarity level. Note that the classifier has approximately equivalent accuracy with 14 very general OTU clusters as it does with 1,282 very specific OTUs clusters.

that are performing those functions. Recent advances in design and inference of complex generative models such as deep belief nets and the many derivatives of topic models may allow us to recover these simple latent factors from the relatively complex communities that we observe.

We may consider the following as a simple generative model for microbial communities: each environment can be viewed as a weighted mixture of (i.e. multinomial distribution over) metabolic functions, where each function is performed by a weighted mixture of species. If all communities in a given data set draw from the same set of metabolic functions and the same set of species, this generative model is known as Latent Dirichlet Allocation (LDA), a popular model from the field of natural language processing [31]. LDA was originally used for automatically extracting conversation topics in the unsupervised semantic analysis of text. If each microbial community is treated as though it were a separate text document in a corpus of documents, the semantic topics in topic modeling are analogous to the metabolic functions or pathways that occur in the communities, and the vocabulary words correspond to OTUs. LDA is a purely generative model; it seeks only to model the distribution of the observed data, $P(D)$, rather than learning to predict class labels based on the data, $P(L|D)$. For the purposes of classification we would of course need to incorporate some discriminative learning into the model. The simplest approach is to "piggy-back" a generic classifier such as RF on top of LDA, using the distribution over latent functions in each community resulting from LDA inference as input features instead of, or in addition to, the raw OTU counts. This approach was used for text classification in the original chapter from Blei et al, and has the potential to work well when the differences between our classification categories are the most important determinant of the mixing proportions of latent functions.

A more direct and more powerful approach is to learn explicitly the joint distribution over category labels and data, $P(L, D)$. This has the potential to combine the strengths of generative and discriminative learning. Several such supervised versions of LDA have been developed, such as Multi-Conditional Learning (MCL) [59] and Supervised LDA (SLDA) [60]. SLDA is the most general, being applicable to many types of response variable including categorical labels (classification) and real-valued labels (regression). MCL is restricted to classification tasks, but was shown by the authors above to perform well in a large variety of text classification problems.

To encourage the evaluation of these types of generative and hybrid (i.e. generative and discriminative) models in future research on microbiota analysis, we show evidence that the latent mixtures over OTUs recovered by classical LDA are indeed related to the category labels in the FSH benchmark. In Figure 6 we show the test set error of an RF classifier using as features the per-community OTU mixtures learned with LDA, plotted against the log-likelihood, given the inferred topic model, of the entire OTU abundance matrix. Each data point was obtained by choosing random values for the LDA model's hyperparameters α, η, uniformly from the interval [0.1,0.5], and then performing collapsed Gibbs sampling to infer a topic model with 25 topics using the *lda* package for *R* [61]. An increase in quality of the fit obtained by the model, as measured by the corresponding log likelihood, is clearly a general indication of an increase in quality of the inferred features as predictors of the class labels (Pearson correlation coefficient = -.51, p-value = 2×10^{-16}).

We also included a simple multinomial naïve Bayes (MNB) classifier [62] in our comparison of classifiers, as shown in Table 2. MNB is the equivalent of a labeled topic model where each class has one topic (mixture of OTUs) that is shared by all of its samples,

and where we learn the topics' mixture components conditioned on the class labels. Our implementation of MNB includes a small prior count for each OTU in each class to act as a smoothing constant, and we chose the value of this constant that minimized cross-validation error within the training set. Although this is a very simple model, it has performance competitive with RF. When compared with the RF, NSC, ENET, and SVM classifiers, MNB achieved the second best mean rank.

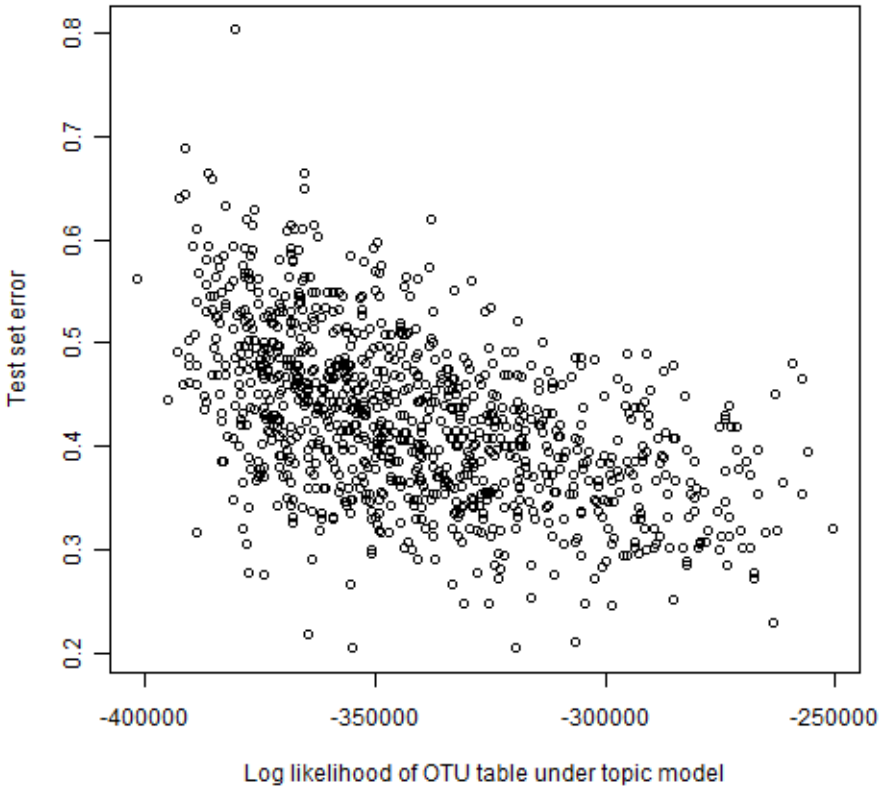

Figure 6. Prediction error versus topic model quality of fit.

Random forests test set error plotted against the log likelihood of the data given a particular topic model. Each data point represents one topic model trained on the entire FSH benchmark data set using randomly chosen values of the topic model's hyperparameters. The latent "topics" recovered by the topic model were then fed into the RF classifier as the only input features. While the error rates here are no better than those in Figure 4 or Table 2, the correlation between the generative log likelihood and the discriminative ability of the derived latent features (topics) implies that topic models may be appropriate generative models for microbial communities; the better the topic model does at modeling the data, the more useful the inferred topics are for explaining differences between communities.

2.9 Data augmentation

Existing high-throughput experiments in microbial ecology typically have many fewer samples than observed species-level taxa, making it difficult to model complex interactions

between the taxa. However, as the number of published experiments grows, there are an increasing number of data sets available that can potentially be used as unlabeled data to augment the labeled training data in a given experiment. Even relatively different data sets may still be useful for training generative and hybrid generative/discriminative models if they contain information about similar OTUs, or even about the ways in which OTUs interact.

For our five benchmarks we found that adding noisy replicates of training data [63, 64] tends to be effective at increasing the predictive power of our models. For each of the benchmark data sets, we generated noisy replicates of the training data by adding a small amount of Gaussian noise (with mean zero and variance equal to the average within-sample variance) to the counts of OTUs present in each sample, thresholding the resulting counts at zero to avoid negative abundance values. We added three noisy replicates of the training set to itself, and fed the augmented train set to an RF classifier with 500 trees. In all cases the expected prediction error when using the augmented training set was as good as or better than that of the best un-augmented model, although the differences were on the order of a 1% or 2% decrease in error. Of course we encourage researchers interested in building supervised classifiers to collect as many samples as possible for a training set, but for cases where there is insufficient training data available, we suggest the exploration of augmented training data, both in the form of noisy training sample replicates and in the form of unlabeled samples from related microbial communities, as an important direction for future research. Such multi-sourced experimentation presents its own challenges; it would at least require uniform labeling of samples (metadata). Metadata standardization efforts such as those by the Genomic Standards Consortium [65] will be essential for large-scale multi-sourced data augmentation.

2.10 Concluding remarks

Supervised learning can serve several purposes for researchers who wish to characterize differences between microbiota in different types of communities. In experiments where the true category membership of communities is well known or is easily obtained, sparse classification techniques such as filter methods or the elastic net can be used to identify specific taxa that are highly discriminative of the categories. The random forest classifier may be useful in these cases as well; although it doesn't explicitly perform any dimensionality reduction, it produces a natural ranking of features by their importance in the model, and it tends to have lower expected prediction error than the other models. In other classification tasks such as forensic identification or the early prediction of disease states, supervised classifiers can be used to learn a predictive model that generalizes well to unseen data. For example, as the cost of DNA sequencing continues to decline, it may become possible to perform gut microbiota surveys of all individuals in a diseased population in order to recommend personalized therapy [25]. In such cases where class prediction is the ultimate goal, one should simply choose whatever model gives us the lowest expected prediction error whether or not it performs explicit feature selection.

We presented five benchmark classification tasks containing data from bacterial 16S rRNA-based surveys of various human body habitats. The benchmarks contain classification tasks of varying difficulty, ranging from distinguishing individual humans by their hand microbiota, which can be done with perfect accuracy, to distinguishing different types of skin sites across individuals, on which task the best classifier we evaluated has 26% expected generalization error. We have made available the same benchmarks as a resource for those interested in pursuing novel techniques for microbiota classification.

All of the supervised classifiers that we reviewed have performed well in similar domains such as microarray analysis or text classification, but it is clear from their performance on our benchmarks that some perform better than others in microbiota classification. Random forests was clearly the strongest performer, being tied for first place in all five of the benchmarks. Multinomial naïve Bayes also tended to perform well, suggesting that generative models like supervised latent Dirichlet allocation may be worth exploring. Support vector machines had surprisingly poor performance without filtering, but they seemed to combine well with the BSS/WSS and SVM-RFE filters. The elastic net classifier tended to have noticeably higher expected error than random forests, although it still proved useful for performing feature selection as a preprocessing step for other classifiers. For example, we included a heatmap of the 27 OTUs selected by the elastic net classifier in the FS benchmark. These OTUs allow >99% test accuracy when trained with the random forests classifier, and thus they represent the unique microbial "fingerprint" of each individual.

Future research into approaches that leverage natural structures inherent in the microbial community data is strongly recommended. Examples include performing dimensionality reduction by reducing the phylogenetic specificity of taxonomic clusters, utilizing the naturally hierarchical structure of features provided by phylogenetic trees, using related data sets as unlabeled data to aid in the inference of generative models, and the exploration of generative or hybrid generative/discriminative techniques to recover latent features, such as metabolic functions, that drive the differences in observed taxa across communities. However, existing classifiers perform well for a range of tasks and will be widely useful in human microbiome projects, perhaps, especially, for identifying biomarkers for disease or other physiological conditions.

CHAPTER 3

3 Applications of supervised learning in microbiome studies

3.1 "Global Gut" analysis[3]

In this broad cross-sectional global survey of the human gut microbiota in varied populations, the contribution from this thesis is the use of supervised classification techniques to determine whether we could discriminate human gut microbial communities by various traits of the host, including nationality, Western/non-Western population status, breastfed versus non-breastfed (for infants). We also used feature selection to identify which genes and species-level sequence clusters were driving the differences between these groups. We determined that we can in fact discriminate adults by their Western/non-Western population status, and even by their geographical region or nationality by their gut microbiota. This study included marker-gene surveys (16S rRNA) of 524 individuals from 147 families, as well as whole-genome shotgun metagenomic surveys of 110 of these individuals. Three populations were surveyed: Malawians (Africa), Venezuelan Amerindians (South America), and the USA (North America).

3.1.1 Contributions from this thesis

We used Random Forests, a supervised machine learning technique [6], and the 16S rRNA datasets obtained from all 524 individuals to identify bacterial species-level operational taxonomic units (OTUs) that identify differences in fecal community composition in children and adults within and between the 3 populations. The purpose of a

[3] From Yatsunenko T, Rey F, Manary M, Trehan I, Dominguez-Bello MG, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Lozupone C, Lauber C, Clemente JC, Knights D, Knight R, Gordon JI. "Human gut microbiome differentiation viewed across cultures, ages and families" (in review, Nature).

classifier such as Random Forests is to learn a function that maps a set of input values or predictors (here, relative OTU abundances in a community) to a discrete output value (here, USA versus non-USA microbiota). Random Forests is a particularly powerful classifier that can exploit non-linear relationships and complex dependencies between OTUs. The measure of the method's success is its ability to correctly classify unseen samples, estimated by training it on a subset of samples, and using it to classify the remaining samples (cross-validation). The cross-validation error is compared to the baseline error that would be achieved by always guessing the most common category. As an added benefit, Random Forests assigns an importance score to each OTU by estimating the increase in error caused by removing that OTU from the set of predictors. In our analysis, we considered an OTU to be highly predictive if its importance score was at least 0.001; all error estimates and OTU importance scores were averaged over 100 even rarefactions of the sample communities in order to control for sequencing effort. For adults, Random Forests revealed distinct community signatures for Western (USA) and non-Western individuals (baseline error=0.286, cross-validation error=0.020 ± 0.004, 64 highly predictive OTUs). Of the 64 highly predictive OTUs, 58 were over-represented in non-USA adults, and 44 of the 58 were assigned to the genus Prevotella or family Prevotellaceae. Malawians and Amerindians could also be distinguished from each other, although the difference was less extreme than the USA versus non-USA comparison (baseline error=0.407, cross-validation error=0.089 ± 0.027, 27 highly predictive OTUs). There were only small discernable differences between infants in the above comparisons, and between adults living in the two Amerindian villages (cross-validation error greater than or equal to half of baseline error in all cases). Thus, a Western (USA) lifestyle appears to affect the bacterial component of the gut microbiota significantly, although this influence is not detectable against the high degree of variability observed in infants and children. Although the Prevotella were the

most discriminatory lineages, removing the entire family of Prevotellacae increased the classification error only slightly, all 20 of the non-Prevotellaceae OTUs are still predictive, and the average decrease in predictive accuracy when they are removed is <0.1%. Thus, as in the case of the Bifidobacteria, the Prevotellaceae provide a major component of the effect we report, but by no means all of the effect.

Confirming the importance of Prevotella as a discriminatory taxon, a recent study also showed that abundance of this genus was present in higher in the fecal microbiota of children living in West Africa (Burkina Faso) compared to children living in Europe (Italy) [66]. Additionally, a member of this genus is one of three bacterial species that, in European adults, distinguishes strongly among three clusters, or enterotypes, of gut microbiota configurations that are claimed to be reproducible across Western adult populations [67]. Therefore, we asked whether the fecal microbiota of infants and adults in each of our three geographically and distinct populations fell into natural discrete clusters. We did not find evidence for discrete clustering, but rather for continuous variation driven in adults by a trade-off between Prevotella and Bacteroides, as previously observed [68]. Although Western and non-Western populations tended to occupy the Bacteroides-rich and Prevotella-rich ends of the gradient, respectively, truncated sections of the gradient were reproduced in each of the three sub-populations we studied. Including infants introduces a new, strongly supported gradient driven by Bifidobacteria, generally orthogonal to the Bacteroides/Prevotella gradient. Clustering of sub-populations of increasing minimum age indicates that adult cluster membership is generally consistent, but that children between 0.6 years and 1 year of age may be clustered with adults or with younger children, depending on whether the younger children are included in the analysis. Therefore, our analysis supports the idea that there is a process of differentiation of an infant community

into adult communities that occurs via a maturation process which is consistent across cultures.

3.2 Long term dietary patterns shape gut microbial enterotypes[4]

We combined a long-term diet inventory with metagenomic DNA sequencing of 98 adult human subjects to determine the long-term effects of diet on the human gut microbiota. The contribution from this thesis was the finding that the gut microbiota of the individuals surveyed fell on a gradient along which the relative abundance of the genera Bacteroides and Prevotella decrease and increase, respectively.

3.2.1 Contributions from this thesis

We performed clustering by partitioning around medoids (PAM) [69] using Jensen-Shannon divergence (JSD) of the normalized genus counts. Weighted UniFrac distance, Euclidean distance and Bray-Curtis distance of the normalized genus counts were also compared. The optimal number of clusters was chosen by the maximum average silhouette width, known as the silhouette coefficient (SC) [70]. The quality of those clusters was assessed by the same measure, following the accepted interpretation that SC values above 0.5 indicate a reasonable clustering structure.

Application of published enterotype clustering methodology. To reproduce a previously published enterotype clustering methodology [67], we performed clustering by PAM using the square root of the Jensen-Shannon divergence, and chose the number of clusters by the Caliński-Harabasz (CH) index of the relative clustering quality as defined in the original publication of the method [71]. The CH index indicated that three clusters were

[4] From: Wu GD, Chen J, Hoffmann C, Bittinger K, Chen Y, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R, Sinha R, Gilroy E, Gupta K, Baldassano R, Nessel L, Li H, Bushman FD, Lewis JD. "Long term dietary patterns shape gut microbial enterotypes". Science 2011 Oct 7;334(6052):105-8.

preferred, but the quality score for three clusters (SC=0.17) indicated no substantial structure. We also applied the CH Index to clustering using several alternative distance measures (Bray-Curtis, Euclidean, Jensen-Shannon, weighted UniFrac, and weighted normalized UniFrac). Interestingly, in all but one case (weighted UniFrac) the CH index chose three as the optimal number of clusters, even though the silhouette scores were substantially higher for two clusters. No reasonable support (SC $\geq$ .5) for three clusters was found using any distance measure.

Prevotella-Bacteroides gradient analysis. The enterotype clustering is driven primarily by the ratio of the two dominant genera, Prevotella to Bacteroides; this ratio defines a clear gradient across the putative COMBO enterotypes, emphasizing that the boundary between enterotypes is not sharply defined. When we removed these genera, the structure was undetectable (17 clusters, SC=0.115). Also, these genera compose between 12% and 83% of the relative abundance in the communities (mean +/- s.d. = 0.46 +/- 0.17), and the only distance measures that produced reasonable support for clustering, JSD and Euclidean distance, are measures that emphasize differences in the largest components of a distribution.

3.3 Diet Drives Convergence in Gut Microbiome Functions Across Mammalian Phylogeny and Within Humans[5]

By characterizing the functional structure (via whole-genome shotgun metagenomics) and phylogenetic structure (via 16S rRNA marker gene sequencing) of 33 mammals and 18 humans, we determined that the convergence of the gut microbiota to similar states across varied mammalian lineages is most highly driven by the diet of those species. The

[5] From: Muegge BD, Kuczynski J, Knights D, Clemente JC, González A, Fontana L, Henrissat B, Knight R, Gordon JI. (2011). "Diet Drives Convergence in Gut Microbiome Functions Across Mammalian Phylogeny and Within Humans." Science 2011, 332(6032):970-974.

contribution of this thesis was the first demonstration that we can predict the functional structure (what genes are present) in a community directly from its phylogenetic structure (what species are present).

3.3.1 Contributions from this thesis

Prediction of Community Functional Profiles from Species Assemblage Data using a Nearest-Neighbor Model. As noted in the main text, the strong correlation between bacterial 16S rRNA and functional profiles made us wonder if the functional configuration of a microbiome could be predicted from its 16S rRNA sequences. To test this idea, we developed a nearest-neighbor model. For a given sample, we predicted its functional composition to be the same as that sample's nearest neighbor (using the weighted UniFrac distance comparison of 16S rRNA data). To assess the quality and significance of these predictions, we compared the average root mean squared error (RMSE) of our model to the average RMSE for one million Monte Carlo trials where each sample's nearest neighbor was chosen at random from the remaining samples. The UniFrac nearest neighbor generated a significantly better functional prediction than a random neighbor for all four types of functional; for KOs, E.C.s, peptidases, and CAZymes, no permutation in the one million trials had a lower RMSE than the UniFrac prediction ($p=0$). Using the unweighted UniFrac distances also led to predicted functional profiles that were significantly better than would be expected by chance (KOs, $p=0$; E.C.s, $p=0$; proteases, $p=0.000252$; CAZymes, $p=0$).

3.4 The impact of a consortium of fermented milk strains on the human gut microbiome: a study involving monozygotic twins and gnotobiotic mice[6]

In this study of the effects of yogurt consumption on the human and mouse gut microbiota, the contribution of this thesis was the use of supervised classification techniques to identify the set of genes that were highly discriminative of the pre/post consumption state in the mouse gut.

3.4.1 Contributions from this thesis

KEGG categories, ECs, and Pathways were all able to predict Pre-/Post- treatment status with low estimated generalization error (KEGG categories: 6.7%, ECs: 13.3%, Pathways: 10.0%). In all cases these generalization error rates were less than half of the baseline error rate (that achieved by always predicting the largest category) of 33%. There were 11 predictive/5 highly predictive KEGG categories, 35 moderately predictive ECs, and 27 predictive/4 highly predictive pathways, shown in Table 4.

To find KEGG categories, ECs, or pathways that were significantly differentiated across treatment states, we used the Random Forests classifier as described previously [1]. Mouse samples were divided into 10 Pre-treatment samples and 20 Post-treatment (21 or 28 days post-treatment) samples. To estimate the generalization error of the classifier we used leave-one-out cross-validation, in which the group for each sample was predicted by a classifier trained on the other 29 samples. Training was done using default settings for the

---

[6] McNulty NP, Yatsunenko T, Hsiao A, Faith JJ, Muegge BD, Goodman AL, Henrissat B, Oozeer R, Cools-Portier S, Gobert G, Chervaux C, Knights D, Lozupone CA, Knight R, Duncan AE, Bain JR, Muehlbauer MJ, Newgard CB, Heath AC, Gordon JI. "The impact of a consortium of fermented milk strains on the human gut microbiome: a study involving monozygotic twins and gnotobiotic mice". **Sci Transl Med**. 2011 Oct 26;3(106):106ra106.

randomForest package in R. The predictiveness of each feature was estimated by calculating the mean increase in estimated generalization error when the values of that feature were permuted at random. Features whose removal caused an average error increase of at least .1% were labeled as predictive; those with an increase of at least 1% were labeled as highly predictive.

KEGG categories	KEGG Pathways	KEGG Enzyme Commision #s
AMINO ACID	Alanine__aspartate_and_glutamate_metabolism	EC1.1.1.103
METABOLISM	Alzheimers_disease	EC1.1.1.40
CARBOHYDRATE	Arachidonic_acid_metabolism	EC1.14.13.3
METABOLISM	Arginine_and_proline_metabolism	EC1.18.1.1
CELL GROWTH	Atrazine_degradation	EC1.3.99.5
AND DEATH	Benzoate_degradation_via_hydroxylation	EC1.4.1.1
CELL MOTILITY	Carbazole_degradation	EC1.4.1.16
AND SECRETION	Carbon_fixation_in_photosynthetic_organisms	EC1.8.4.11
ENVIRONMENTAL	Fatty_acid_metabolism	EC2.1.1.113
ADAPTATION	Flavone_and_flavonol_biosynthesis	EC2.6.1.1
FUNCTION	Fructose_and_mannose_metabolism	EC2.6.1.9
UNKNOWN	Inositol_phosphate_metabolism	EC3.1.1.11
GERMINATION	Isoquinoline_alkaloid_biosynthesis	EC3.1.26.-
LIPID	Lysosome	EC3.1.3.6
METABOLISM	Nitrogen_metabolism	EC3.1.6.6
METABOLISM OF	Novobiocin_biosynthesis	EC3.2.1.18
OTHER AMINO	Other_glycan_degradation	EC3.2.1.89
ACIDS	Pentose_and_glucuronate_interconversions	EC3.4.24.75
PORES ION	Phenylalanine_metabolism	EC3.5.-.-
CHANNELS	Prion_diseases	EC3.5.1.10
SIGNALING	Pyruvate_metabolism	EC3.5.4.-
MOLECULES AND	Selenoamino_acid_metabolism	EC3.5.4.25
INTERACTION	Sphingolipid_metabolism	EC3.5.99.2
SPORULATION	Starch_and_sucrose_metabolism	EC4.1.3.-
TRANSCRIPTION	Steroid_hormone_biosynthesis	EC4.1.3.39
TRANSCRIPTION	Streptomycin_biosynthesis	EC4.1.99.12
RELATED	Styrene_degradation	EC4.2.1.44
PROTEINS	Sulfur_metabolism	EC4.2.1.47
TRANSLATION	Taurine_and_hypotaurine_metabolism	EC5.2.1.8
TRANSPORT AND	Translation_factors	EC5.3.1.14
CATABOLISM	beta-Alanine_metabolism	EC5.3.1.5
		EC6.1.1.3
		EC6.1.1.4
		EC6.2.1.34
		EC6.3.2.-

Table 4. Highly predictive KEGG features for discriminating pre-/post-yogurt mouse gut communities.

Those with (random forests' mean decrease in accuracy > 1%) are considered highly predictive.

CHAPTER 4

4 Bayesian community-wide microbial source tracking[7]

Contamination is a critical issue in high-throughput metagenomic studies, yet progress towards a comprehensive solution has been limited. We present SourceTracker, a Bayesian approach to estimating the proportion of a novel community that comes from a set of source environments. We apply SourceTracker to new microbial surveys from neonatal intensive care units (NICUs), offices, and molecular biology laboratories, and provide a database of known contaminants for future testing.

4.1 Background

Advances in sequencing technology and informatics, including the MIxS (Minimum Information about any (x) Sequence) metadata standards, are producing an exponential increase in data acquisition and integration. These advances are revolutionizing our understanding of the roles microbes play in health and disease, biogeochemical cycling, etc. Although considerable attention has been paid to reducing sources of error from PCR [72] and sequencing [35], sample contamination has been relatively unstudied. Preparing contaminant-free DNA is challenging, and the sensitivity of PCR and whole-genome amplification methods means that even trace contamination can become a serious issue [73]. Ideally, computational methods could identify both the source and quantity of contamination, and could help prevent future instances. Furthermore, accurately

[7] From: Knights D, Kuczynski J, Charlson E, Zaneveld J, Collman RG, Bushman FD, Knight R, Kelley ST. (2011). Bayesian community-wide microbial source tracking. Nat Methods. 2011 Jul 17.

estimating the proportion of contamination from a given source environment would have far-reaching applications in source tracking for forensics, pollution, public health, etc.

4.2 Overview

We have developed SourceTracker, a Bayesian approach to identifying sources and proportions of contamination in marker-gene and functional metagenomics studies. Our approach models contamination as the mixture of entire source communities into a sink community, where the mixing proportions are unknown. Previous approaches to microbial source tracking (MST) have focused on detection of fecal contamination in water [12, 74, 75], limited to detection of predetermined indicator species and custom-tailored biomarkers from source communities. One notable exception [10] uses community structure to measure similarity between sink samples and potential source environments. Other prior work uses data-driven identification of indicator species, but lacks a probabilistic framework [11]. SourceTracker's distinguishing features are its direct estimation of source proportions, and its Bayesian modeling of uncertainty about known and unknown source environments.

We also present barcoded pyrosequencing datasets of bacterial 16S ribosomal RNA gene sequences covering surface contamination in office buildings, hospitals, and research labs, and reagents used for metagenomics studies (data collection described in Online Methods). Using SourceTracker, we compared these data to published datasets from environments likely to be sources of indoor contaminants, namely human skin, oral cavities, and feces [20], and temperate soils [76]. We treated these natural environments as sources contributing organisms to the indoor sink environments through natural migration (as with office samples) or inadvertent contamination (as with no-template PCR controls) (schematic in Figure 7).

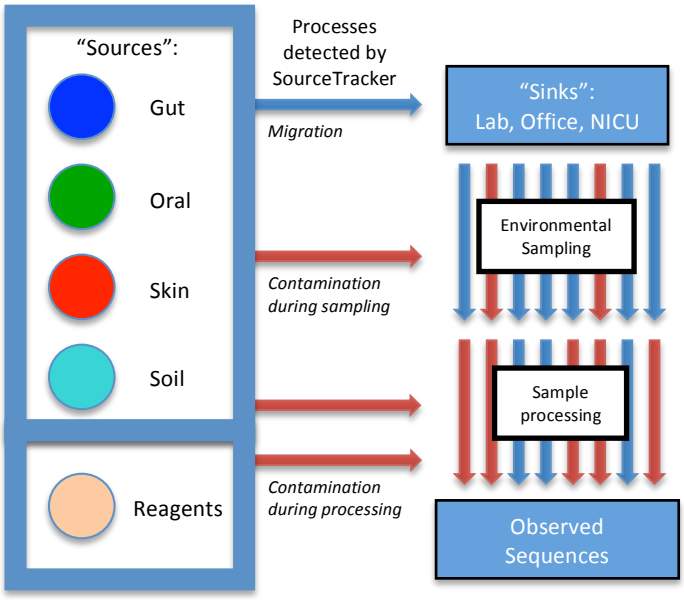

Figure 7. Schematic of SourceTracker analysis.

SourceTracker allows tracking of source environments that have contributed organisms to sink samples. Transfer of microbes may occur in nature prior to sampling (migration), or during the sampling and processing steps (contamination). To demonstrate the utility of SourceTracker, we addressed the question of which sources may commonly contribute to the microbial communities on the surfaces of indoor environments including laboratories, offices, and NICUs. Several of the sink environments characterized (PCR water, laboratory benches) are themselves potential sources of contamination, and they contribute to the library of potential sources that we envision tracking with SourceTracker in future environmental samples.

Although qualitative assessment of source and sink similarities can be performed by visualizing UniFrac distances [57] (Figure 8), this cannot tell us the proportion of each sink sample (e.g., a cotton swab) comprising taxa from a known source environment (e.g., soil). The problem would be trivial if source and sink environments shared no taxa, but usually some taxa are shared. Source tracking methods must therefore leverage potentially useful information contained in the abundance of species with low or moderate source environment endemicity.

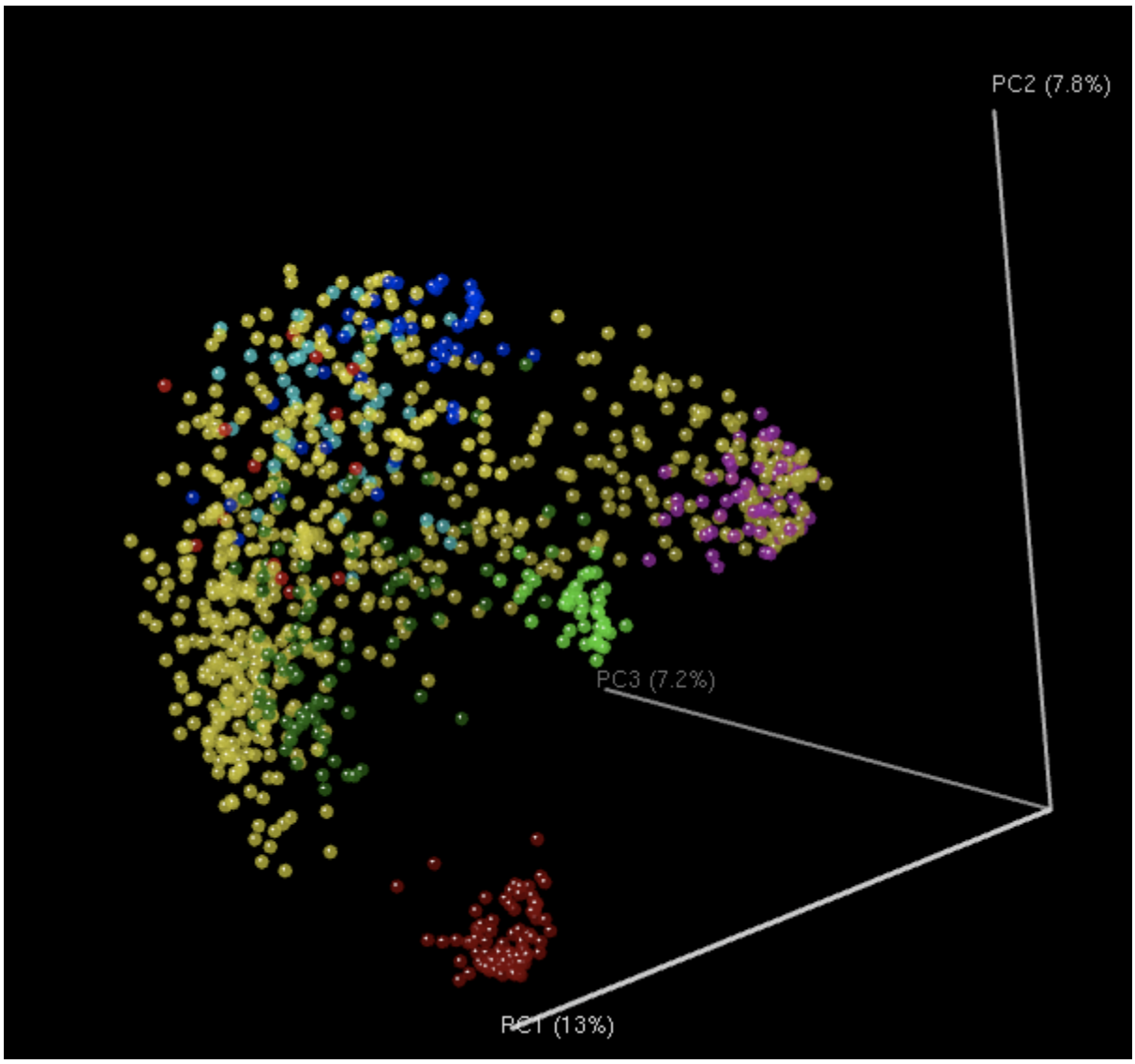

Figure 8. Principal Coordinates analysis of Source and Sink samples.

dark green; Soil samples: dark red (bottom of figure); Whole body samples as follows: External ear canal: dark blue; Gut: light green (center of figure); Hair: red (center left); Nostril: light blue; Oral: purple; Skin: yellow. The first 3 principal coordinates are shown, rotated to display most clearly the variation among groups in 2 dimensions.

Previous work uses probabilistic indicator species for naïve Bayes estimation [75]. Although naïve Bayes actually estimates the probability that each source generated the entire sink sample, these probabilities can sometimes act as proxies for the proportions of

the sink contributed by each source. We compared the accuracies of naïve Bayes and SourceTracker as we varied the distributions of taxa in two simulated source environments from perfectly identical to perfectly non-overlapping, with and without an Unknown source included (Figure 9). Naïve Bayes was accurate when disambiguation is easy, but inaccurate elsewhere. SourceTracker performed well even when disambiguation is difficult ($R^2 \geq .8$, Jensen-Shannon divergence ≥ 0.05). We also evaluated the accuracy of a simple linear regression model with no "Unknown" component. Linear regression generally performed better than naïve Bayes, but worse than SourceTracker when there was an Unknown source. SourceTracker outperforms these methods because it allows uncertainty in the source and sink distributions, and because it explicitly models a sink sample as a mixture of sources.

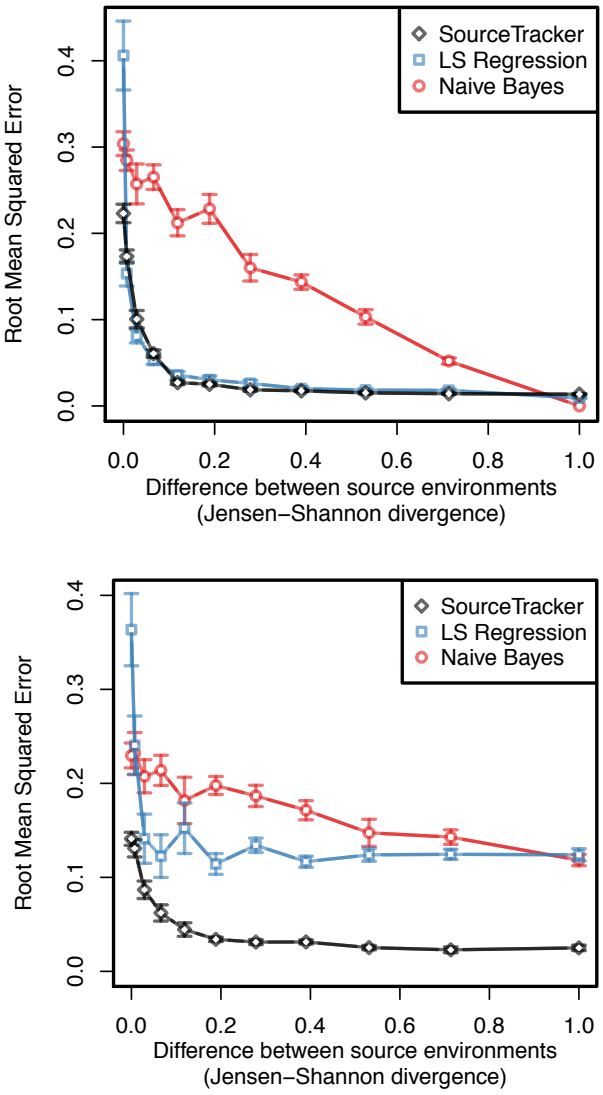

Figure 9. Performance of SourceTracker and other models on simulated data.

We varied the proportion of overlap between two simulated source communities. In the top panel, there are two known sources; in the bottom panel, there is 25% of a randomly generated Unknown community.

4.3 The SourceTracker model

The Bayesian approach requires consideration of all possible assignments of the test sample sequences to the different source environments, but direct exploration is

intractable. Fortunately, we can explore this joint distribution using Gibbs sampling, a technique widely used in the exploration of complex posterior distributions in applications like topic modeling [31]. Community-wide source tracking is analogous to inferring the mixing proportions of conversation topics in a test document, except that the source environment distributions over taxa (topic distributions over words) are known from the training data, and each test sample may contain taxa from an unknown, uncharacterized source. The application of Gibbs sampling to topic modeling has been discussed in detail previously [77].

SourceTracker considers each sink sample x as a set of n sequences mapped to taxa, where each sequence can be assigned to any one of the source environments $v \in \{1..V\}$, including an Unknown source. These assignments are treated as hidden variables, denoted $z_{i=1..n} \in \{1..V\}$. To perform Gibbs sampling, we initialize z with random source environment assignments, and then iteratively re-assign each sequence based on the conditional distribution:

$$P(\mathbf{z}_i = v \mid \mathbf{z}^{\neg i}, \mathbf{x}) \propto P(\mathbf{x}_i \mid v) \times P(v \mid \mathbf{x}^{\neg i}) = \left(\frac{m_{x_i v} + \alpha}{m_{\cdot v} + \alpha m_{\cdot v}} \right) \times \left(\frac{n_v^{\neg i} + \beta}{n - 1 + \beta V} \right),$$

where m_{tv} is the number of training sequences from taxon t in environment v, n_v is the number of test sequences currently assigned to environment v, and $\neg i$ excludes the i^{th} sequence. The first fraction gives the posterior distribution over taxa in the source environment; the second gives the posterior distribution over source environments in the test sample. Both are Dirichlet distributions, and Gibbs sampling allows us to integrate over their uncertainty. The Dirichlet parameters, a and b, act as imaginary prior counts that smooth the distributions for low-coverage source and sink samples, respectively. They also allow Unknown source assignments to accumulate when part of a sink sample is unlike

any of the known sources. By inferring source proportions for multiple sink samples simultaneously, we can allow them to share an Unknown source. We could also include several Unknown sources. Full details and an overview of Gibbs sampling are provided in our Online Methods.

4.4 Applications and Validation

For each of our indoor sink environments, we used SourceTracker to estimate the proportion of bacteria from Gut, Oral, Skin, Soil, and Unknown (i.e., one or more sources absent from the training data) (Figures 10, Figure 11,Figure 12). In general, wet-lab surface communities tended to be composed mainly of bacteria from Skin and Unknown, with the exception of PCR water, which was generally more similar to Gut. NICU and office communities were dominated by Skin bacteria, except for two Arizona samples dominated by Soil bacteria and several telephone samples dominated by Oral bacteria. From these results we can also determine the most common contaminating taxa (Figure 13).

For low-coverage sink samples, or when source environments lack a "core" set of taxa, SourceTracker will report high variability in the proportion estimates (Figure 10). In some data sets, variation within each source environment (the "non-core" taxa) might be accounted for by using phylogenetic information, by automatically identifying distinct niches within the broader source environment, by modeling post-mixture population dynamics, or by modeling potential biases inherent in the DNA extraction procedures used; these are important directions for future work. SourceTracker also assumes that an environment cannot be both a source and a sink, and we recommend research into bi-directional models.

SourceTracker can also be used to detect low-level contamination, with sensitivity adjusted by the prior parameter *b*. For simulations with 1% and 5% contamination, SourceTracker achieved nearly perfect specificity for a wide range of sensitivities, demonstrating that it is not restricted to low-biomass sink environments where contamination rates are likely to be higher (area under the receiver operating characteristic curve = .971 for 1%, .989 for 5%; Figure 14).

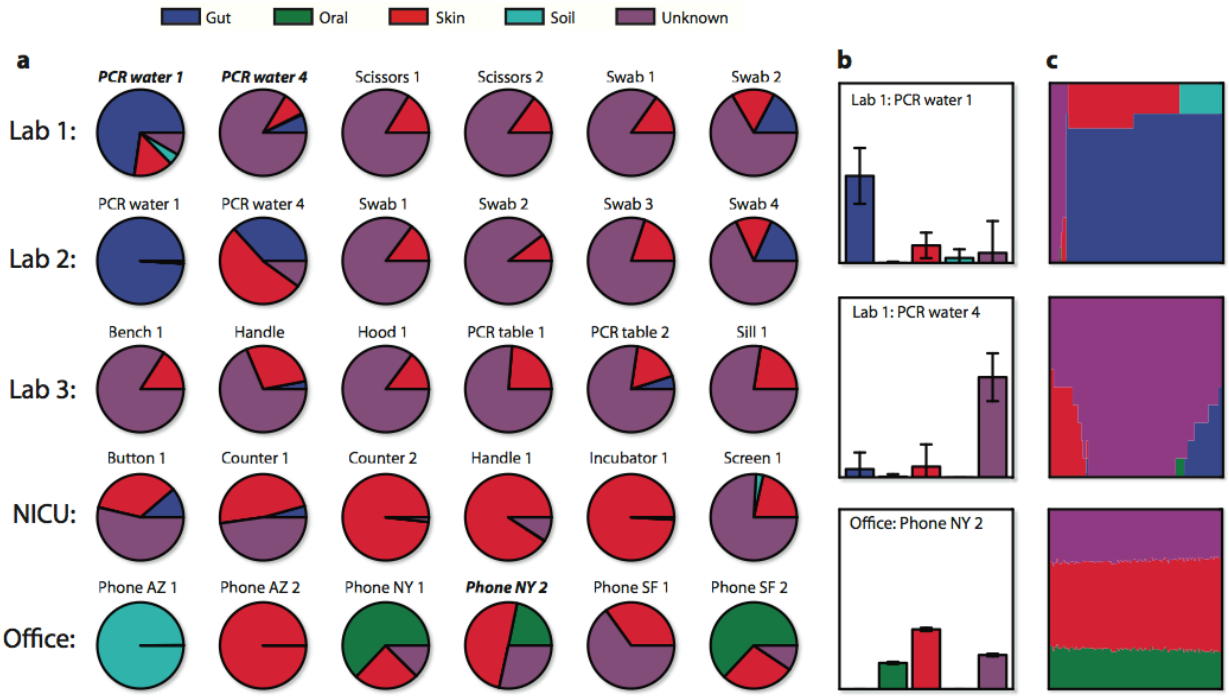

Figure 10. SourceTracker proportion estimates for a subset of sink samples.

Source environment proportions were estimated using SourceTracker and 45 training samples from each source environment. (a) Pie charts of the mean proportions for 100 draws from Gibbs sampling. (b) Bar charts for three samples including standard deviations of the proportion estimates. (c) Direct visualization of 100 Gibbs draws for the samples in (b); each column shows the mixture from one draw, with columns sorted by the most prevalent source. The first sample, Lab 1: PCR water 1, shows several possible mixtures: all Unknown; Gut and Skin (most common); and Gut and Soil. The second sample shows poor disambiguation between Gut, Skin, and Unknown. Most mixtures were stable like the third sample; the first two were chosen for demonstrative purposes.

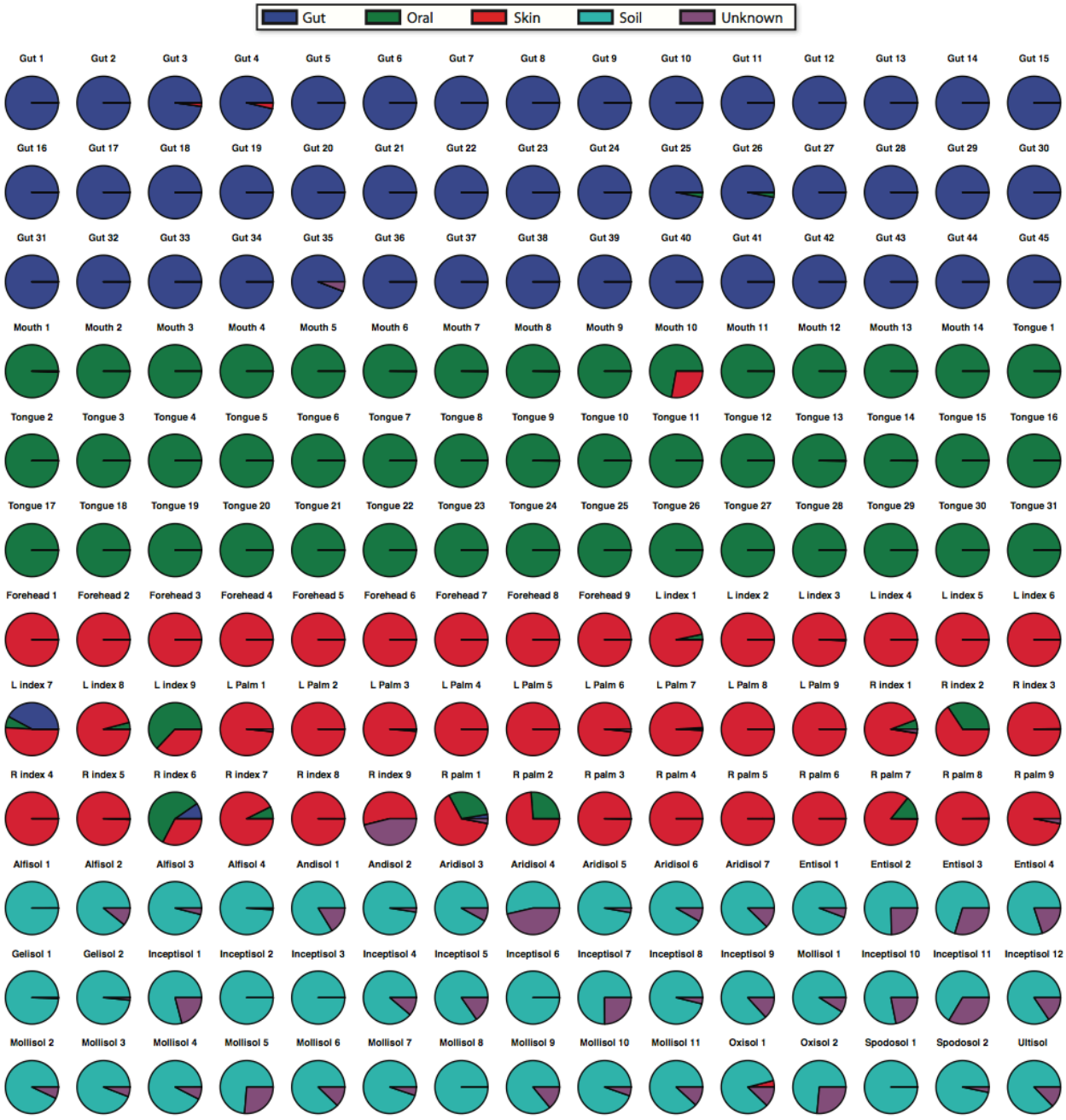

Figure 11. Source proportion estimates were predicted using a leave-one-out approach.

For a given sample, that sample was removed from the training process (estimation of the distributions over taxa in each source environment), and then treated as a single sink sample for estimation. The first three rows are Gut, the next three Oral, the next three Skin, and the last three Soil. The higher prevalence of Unknown bacteria in the Soil samples is an indication that the soil training set has less of a "core" microbiome than the other source training sets.

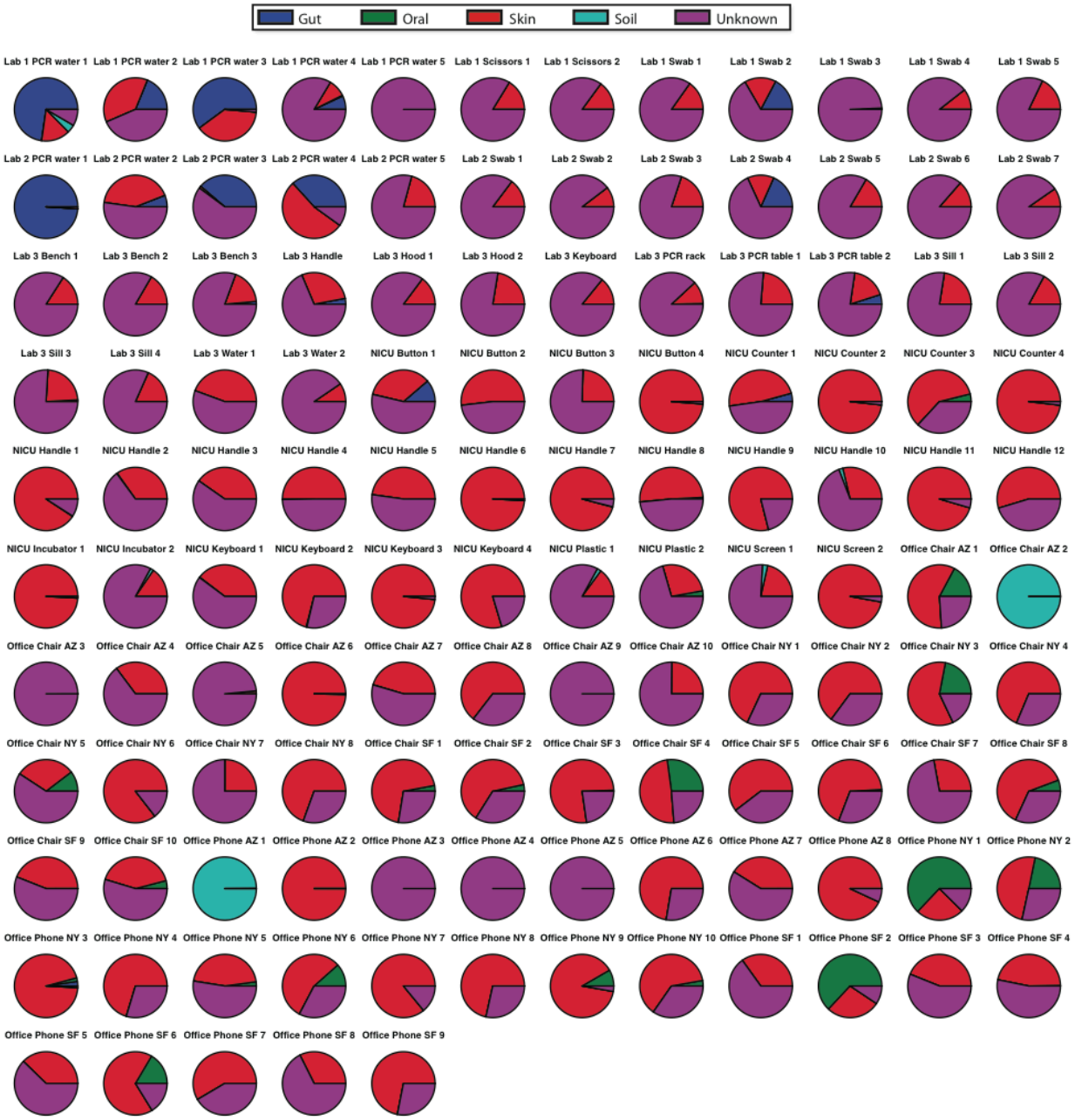

Figure 12. Estimated composition of all sink samples using SourceTracker.

Source environment proportions were estimated using SourceTracker and 45 training samples from each source environment (Supplementary Table 2). The pie charts show the mean proportions for 100 draws from Gibbs sampling.

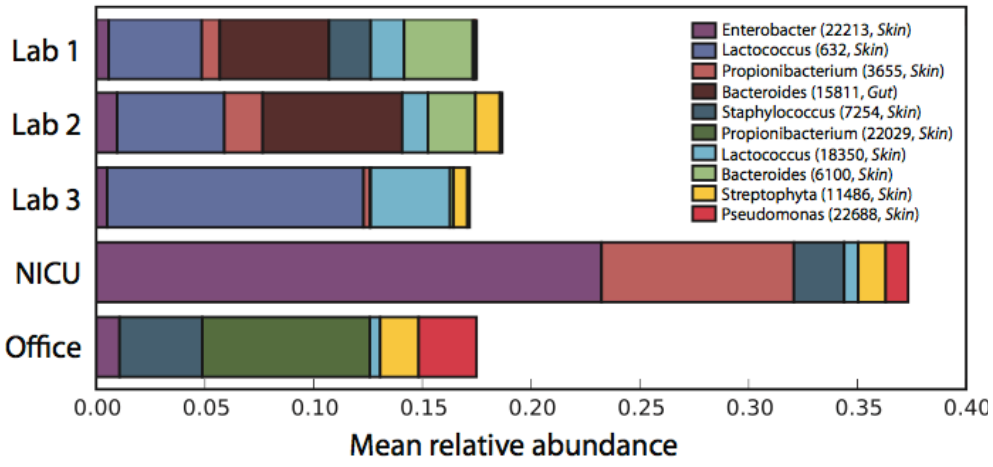

Figure 13. Relative abundance of common contaminating operational taxonomic units (OTUs).

For all sink sequences assigned to a known source environment (Gut, Oral, Skin, or Soil) by SourceTracker, these ten OTUs had the highest average relative abundance across sink environments. Note that the OTU classified as Enterobacter, a lineage commonly seen in the gut, was more prevalent in the Skin training samples than the Gut training samples.

4.5 Conclusion

Based on our results, simple analytical steps can be suggested for tracking sources and assessing contamination in newly acquired data sets. Although source-tracking estimates are limited by the comprehensiveness of the source environments used for training, large-scale projects such as the Earth Microbiome Project will dramatically expand the availability of such resources. SourceTracker is applicable not only to source tracking and forensic analysis in a wide variety of microbial community surveys (e.g., "where did this biofilm come from?"), but also to shotgun metagenomics and other population genetics data. We have made our implementation of SourceTracker available as an R package (http://sourcetracker.sf.net), and we advocate automated tests of deposited data to screen samples that may be contaminated prior to deposition.

4.6 Online Methods

Data collection. We collected the Office samples from surfaces in 54 offices in three different office buildings (18 per building) located in New York, NY; San Francisco, CA; and Tucson, AZ, respectively (Hewitt, K.M., Gerba, C.P., Maxwell, S.L. & S.T.K., unpub. data). In each office, we sampled the same two surfaces, phone and chair, by swabbing approximately 13 cm^2 with dual tip sterile cotton swabs (BBL CultureSwab™, catalog # 220135). Phone and chairs had already been determined by culture-based methods to be the most contaminated surfaces in these offices (unpub. data). We also collected samples from surfaces in two different large Level three Neonatal Intensive Care Units (NICUs) in San Diego, CA using the same methods. After sampling, we stored swabs in sterile-labeled tubes, placed them on ice and shipped them overnight, or drove them directly to the lab for DNA extraction.

For the Lab 1 and Lab 2 data sets, we cut sterile nylon-flocked swabs (Copan) and swabs of sterile scissors into MoBio 0.7 mm garnet bead tubes (Mo Bio Laboratories) using autoclaved and flamed scissors in a biosafety cabinet, placed them at -80°C within 1 hour, and stored them for <1 week prior to DNA extraction.

For the Lab 3 data set, we used sterile nylon-flocked swabs (Copan) to sample indoor surfaces including desktops, lab benches, windowsills, a keyboard, and a door handle over a three-month period from January-March 2010 in Philadelphia, PA. We cut swabs into MoBio 0.7 mm garnet bead tubes (Mo Bio Laboratories) using autoclaved and flamed scissors in a biosafety cabinet, placed them at -80°C within 1 hour, and stored them for <1 week prior to DNA extraction.

DNA extraction, PCR, and pyrosequencing. For the Office and NICU samples, we removed the cotton from the swab using a flame-sterilized razor blade and deposited the

cotton threads into a lysozyme reaction mixture. The reaction mixture had a total volume of 200 µl and included the following final concentration: 20 M Tris, 2 mM EDTA (pH 8.0), 1.2% P40 detergent, 20 mg ml-1 lysozyme, and 0.2 µm filtered sterile water (Sigma Chemical Co.). We incubated the samples in a 37°C water bath for thirty minutes. Next, we added Proteinase K (DNeasy Tissue Kit, Qiagen Corporation) and AL Buffer (DNeasy Tissue Kit, Qiagen Corporation) to the tubes and gently mixed them. We incubated the samples in a 70°C water bath for 10 min. We subjected all samples to purification using the DNeasy Tissue Kit. Following extraction, we quantified the DNA using a NanoDrop ND-1000 Spectrophotometer (NanoDropTechnologies). PCR barcoded primers and conditions were previously described [19]. PCR purification, dilutions and pyrosequencing (FLX) were all conducted by the core facility at the University of South Carolina (Environmental Genomics Core Facility).

For the Lab 1 and Lab 2 data sets, we extracted genomic DNA from swabs using the QIAamp DNA Stool Minikit (Qiagen) with the following modifications. We added 1500 ul of ASL buffer and 5mM DTT to the nylon tips of frozen swabs. We beadbeat tubes with BioSpec Products Inc. Minibeadbeater-16 for 1 min. and incubated at 95 ˚C for 10 min. We performed the remaining steps as per manufacturer protocol. We performed PCR amplification of 16S rRNA genes using the V1V2 primers and conditions described in Wu et al. [78] in duplicate. We quantified purified amplicons using Quant-iT PicoGreen kit (Invitrogen) and pooled them in equimolar ratios. We also performed PCR on molecular biology grade water (Sigma) and included it in the pool. We carried out pyrosequencing using primer A and the Titanium amplicon kit on a 454 Life Sciences Genome Sequencer FLX instrument (Roche).

For the Lab 3 data set, we extracted genomic DNA from swabs using the same extraction kit and technique as Lab 1 and 2 above. We performed PCR amplification of 16S rRNA genes using the V1V2 primers and conditions described in Wu et al., 2010. We quantified purified amplicons using Quant-iT PicoGreen kit (Invitrogen) and pooled them in equimolar ratios. We also performed PCR on molecular biology grade water (Sigma) and included it in the pool. We carried out pyrosequencing using primer A and the Titanium amplicon kit on a 454 Life Sciences Genome Sequencer FLX instrument (Roche).

Combined preprocessing of contamination data sets. We processed the DNA sequence data for all source and sink samples in combination using the QIIME pipeline [36]. In order to avoid bias, we selected subsets of the same size (45 samples) from each of the four source environments (Supplementary Table 2). We sequenced samples in multiplex using error-correcting nucleotide barcodes, and we used QIIME to demultiplex the samples and perform quality filtering. We then used flowgram clustering [79] to remove sequencing noise. We clustered similar sequences (≥ 97% similarity) into OTUs with uclust [80], and assigned taxonomic identity to each OTU using the Ribosomal Database Project's taxonomy assignment tool [23]. We aligned representative sequences from each OTU against the greengenes reference 'core set' of 16S rRNA gene sequences (http://greengenes.lbl.gov). We then removed likely chimeric PCR products using Chimera Slayer [81]. We used the remaining aligned sequences to construct a phylogeny relating the sequences, via FastTree [82].

Identification and removal of Chimeras. As noted above, we removed likely chimeric PCR products using Chimera Slayer [81]. Note that we first aligned representative sequences from each OTU to the greengenes core set. Any OTU not aligning to the greengenes core set at > 75% identity to the nearest BLAST hit in the core set was

discarded. These discarded sequences may contain chimeras, as well as other artifacts. However, once completed we also used Chimera Slayer to screen the resulting sequences for chimeras. The number of chimeras removed were: 58 sequences from Lab 1 samples (4%), 105 from Lab 2 (4%), 4208 from Lab 3 (5%), 422 from Office (0.3%), and 1365 from NICU (0.6%).

Principal Coordinates Plots. After randomly selecting 500 sequence reads per sample and dropping low-coverage samples to control for sequencing effort, we used UniFrac [57] to measure the phylogenetic dissimilarity of all samples and performed Principal Coordinates Analysis (PCoA) on the matrix of unweighted UniFrac distances using QIIME [36].

Gibbs sampling overview. To begin the Gibbs sampling procedure we assign each sequence to a random source environment. We assume that these assignments are correct (even though they are random), and tally the current proportions of the source environments in the test sample. We then remove one sequence from the tallies and re-select its source environment assignment, where the probability of selecting each source is proportional to the probability of observing that sequence's taxon in that source, multiplied by the current estimate of the probability of observing that source in the test sample. After the re-assignment, we update the tally for the selected source environment, and repeat the process on another randomly selected sequence. After we have re-assigned all of the sequences many times in this manner, each set of assignments we observe is a representative draw from the distribution over all possible sequence-source assignments. To estimate the variability of this distribution, we can repeat the procedure as many times as we like, and we can report summary statistics for the mixing proportions or even visualize their distributions directly.

Dirichlet prior parameters. A larger value of b causes a smoother posterior distribution over environments in the sink sample. This is valuable when we want to avoid overfitting in sink samples with few sequences. By assigning different relative values of b to each environment, we can also incorporate prior knowledge about the expected distribution of source environments in our sink samples. a represents a prior count of each taxon in each source environment. This allows taxa that are unlikely under the known source environment distributions to accumulate in an Unknown environment during the sampling procedure. In order to simplify the choice of values for a and b, we treat them as prior counts *relative to* the number of sequences in the test sample, rather than absolute prior counts. For all inferences performed in this chapter, we set both a and b to 0.0001. We use a separate and larger value of a (0.1) for the prior counts of each taxon in the Unknown environment, in order to prevent that environment from overfitting each individual test sample. If we had a prior belief that some of the test samples shared the same Unknown environment, we could perform inference on them jointly, and reduce this separate a value accordingly. In practice, we can train the values of these hyperparameters using cross-validation within the source environment samples.

As is typical in Gibbs sampling, we first performed a set of "burn-in" passes (25 passes) through the entire set of sequences in a data sample before drawing a mixture sample from the joint posterior. We also re-started the entire sampling process with new random hidden variable values 100 times, thereby collecting a total of 100 samples from the posterior distribution for each sample. Each iteration on a sink sample with V source environments requires $O(V^2 n)$ operations. Before running Gibbs sampling, we rarefied all samples to an artificial sequence depth of 1,000. We kept any samples whose original sequence depth was less than 1,000 at that lower depth.

Simulations. For the comparison of SourceTracker to naïve Bayes and Linear Regression, we simulated two source environments with varying degrees of overlap in their distribution over taxa by defining two uniform Dirichlet priors over 100 taxa with a concentration of $\alpha = 1$, where half of the taxa are absent ($\alpha = 0$) from one prior and the other half are absent from the other prior. We then generate random deviates from these two Dirichlets and mix them at varying ratios. These form the base distributions for the simulated source and sink training samples. By varying these ratios, we were able to control the degree of overlap between the two multinomials. The "Unknown" source base distribution was generated from a Dirichlet with uniform prior $\alpha = 1$.

For the application of SourceTracker with Gibbs sampling to the detection task, we used all of the Gut and Skin training samples to estimate the multinomial distribution over taxa in each environment. To generate "contaminated" samples, we drew 100 simulated samples from each environment at sequencing depth 1,000 and mixed them together with 1% (or 5%) Skin and 99% (or 95%) Gut. We also generated 100 pure Gut samples at depth 1,000. We then ran SourceTracker as described above to estimate the proportion of Skin taxa in the simulated Gut samples. We used a contamination threshold of one-half of the contamination rate, and varied the Dirichlet parameter b to adjust the sensitivity of the model (higher b means higher sensitivity). For each value of b, with its corresponding level of sensitivity, we measured the specificity of the contamination predictions made by SourceTracker, and plotted the series of values as receiver operating characteristic curves (Figure 14).

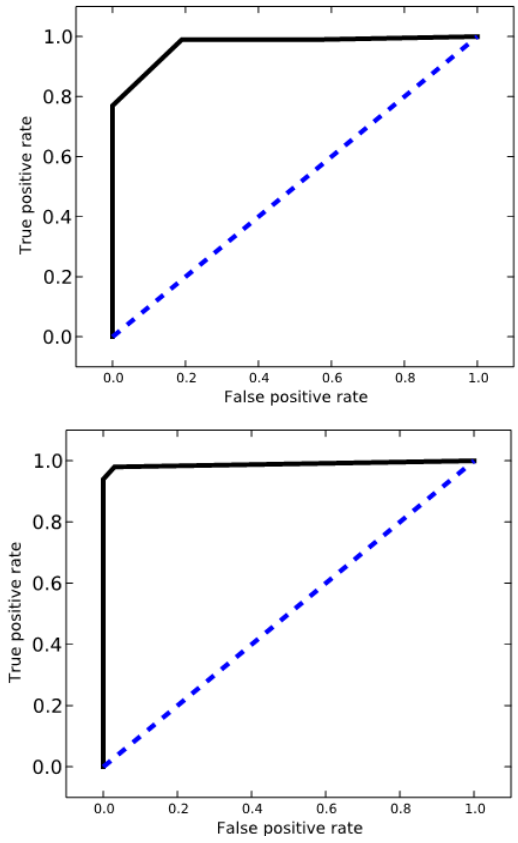

Figure 14. ROC curves for detecting simulated contamination using SourceTracker.

(a) 1% contamination, area under curve (AOC) is 0.971; (b) 5% contamination, AOC is 0.989.

CHAPTER 5

5 Conclusion[8]

Host-associated microbial communities are unique to individuals and can affect host health in a variety of ways. They also correlate with disease states and could potentially be used in forensic settings. Although the need for predictive models of human microbiota is urgent in areas such as personalized medicine, early detection of disease states, forensic identification, source tracking, and detection of contamination, to date there has been little research into novel approaches. Recent advances in DNA sequencing technology now allow us to capture detailed snapshots of microbial communities at dramatically reduced cost. However, traditional statistical inference techniques for assessing community differences and distinguishing host phenotypes for the communities are complicated by high within- and between-subject variation. In this thesis, we suggest turning to machine learning, which has been applied extensively to other high-dimensional domains such as microarray analysis and text classification, and has been demonstrated to be effective for the classification of microbial communities. We discuss key directions for future development and application to harnessing human associated microbial signatures for medical and forensic uses.

5.1 Discovery of microbial signatures

Many human diseases are caused by single species or strains of bacteria, such as tuberculosis (*Mycobacterium tuberculosis*), tetanus (*Clostridium tetani*), and diphtheria (*Corynebacterium diphtheriae*); these specific taxa, along with their associations to host

[8] From: Knights D, Parfrey LW, Zaneveld J, Lozupone C, Knight R. "Human-associated microbial signatures: examining their predictive value". Cell Host & Microbe 2011 Oct 4;10(4):292-6.

phenotypes, are sometimes referred to as biomarkers. Diagnosis and prevention of these types of diseases is relatively simple: if you have the biomarker, you have the disease. Similarly, tracking pathogens and contaminants in environmental samples has traditionally focused on counts of a single species, such as E. coli, or group of species, such as coliforms [12]. In the age of high-throughput DNA sequencing, discovery and verification of individual biomarkers for various host phenotypes is straightforward: collect and sequence enough data from hosts with and without the phenotype, and a classical hypothesis test (e.g. t-test or Mann-Whitney *U* test) will detect differential abundance of the biomarker. But there may be other cases when there is no single biomarker for a phenotype. We know now that host-associated bacterial communities are composed of hundreds or thousands of unique species, and many host phenotypes are associated with shifts in bacterial communities, but not with specific causative agents. For example, let us consider a hypothetical enteric disease state that is associated with concurrent overrepresentation of the phylum Bacteroidetes, the genus *Shigella* and the species *Helicobacter pylori*. We now have a three-way interaction between three different lineages of varying phylogenetic depth. We could refer to this set of interacting biomarkers, and the relationship that they have with the host phenotype, as a *microbial signature*. Such a signature need not be limited to taxonomic characterizations of communities (e.g. surveys of marker genes such as 16S rRNA) but may also include genes or functional categories.

As illustrated in the example above, a microbial signature may be arbitrarily complex, involving simultaneous over- and under-representations of multiple taxa at multiple taxonomic levels. In some cases, the traits that lead to disease may be limited to a single bacterial strain (perhaps one that has acquired virulent factors on a plasmid), while in others these traits may be more phylogenetically conserved, such that treating a whole

genus or family as a feature would be optimal for dimensionality reduction. Given a hypothetical data set containing 1,000 unique species (pragmatically defined as 97% OTUs, or organisms with at least 97% identity in their 16S rRNA sequences), we would have to perform approximately one billion classical hypothesis tests to explore all such interactions at all taxonomic ranks, and controlling the rate of false positives would be next to impossible. Within these complex communities, how can we determine which lineages or genes matter, and at what taxonomic level, for a given host phenotype?

The discovery of such relationships is the goal of supervised learning: we use a set of communities with known phenotype to train a machine learning algorithm; the algorithm identifies discriminative independent variables and produces a *predictive model* which can then be used to predict the phenotype associated with other microbial communities. The machine learning community refers to this approach as "supervised learning", or "supervised classification" (this use of the term "classification" is not to be confused with taxonomic classification of individual sequences or OTUs). Supervised learning is essentially a formalization of the implicit goal of most exploratory scientific research: based on the results of an experiment, we propose a descriptive model (e.g. a linear regression) that we believe will hold true for similar experiments in the future. What distinguishes supervised learning from classical hypothesis testing is that supervised learning deals explicitly with estimating and improving the expected future accuracy of a predictive model at the same time that it is discovering predictive signatures—they are two parts of the same process. There are extensive and varied approaches within machine learning devoted to building predictive models and maximizing their expected accuracy (previously reviewed in the context of microbial community classification [1]).

For simplicity we have focused so far on scenarios involving diagnosis of disease states, but we also envision potential applications in prognosis of treatment response, forensic identification of the host, and detection and sourcing of environmental sample contamination. In the context of these potential applications, we now discuss several remaining challenges in the discovery of predictive microbial signatures.

5.2 Improving discovery with existing biological knowledge

In many ways, studies of the microbiome can be informed by the extensive work that has been done in the closely related area of microarray classification [29], although there are some important distinctions [1]. Both microarrays and high-throughput characterizations of microbial communities such as marker-gene surveys or shotgun metagenomics produce high-dimensional data. However, unlike gene expression data, the low degree of overlap in species among subjects, for example, in the human gut, also leads to very sparse data matrices (i.e. matrices that contain many zeros) in marker gene surveys. The dual challenges of high dimensionality and high sparsity make it hard to identify individual biomarkers. Much of the work on predictive modeling of microarray data has focused on removing noisy or irrelevant independent variables (genes) from the data [29]. In the field of machine learning this process of identifying and discarding noisy independent variables (e.g., taxa or genes) is often referred to as “feature selection”. Feature selection is similar to controlling the Type I error rate for multiple individual hypothesis tests, but the underlying motivation is to reduce the expected error of the model when it classifies novel communities.

Several existing feature selection techniques are helpful for classifying microbial communities [1]. However, it is likely that we can also take advantage of relational or hierarchical structures in the data such as taxonomies, gene ontologies, metabolic

pathways, etc. (Figure 15) to share statistical strength between weakly predictive independent variables. One important consideration is that the abundance of taxa or genes is usually measured in relative terms. In this case the data are compositional, that is, when the relative abundance of one taxon increases, the relative abundance of the rest of the community must necessarily decrease. Consequently, explicit modeling of compositional distributions may be appropriate. One such probability distribution, the Dirichlet, has already been effective for community-wide microbial source tracking [3].

The hardest part of detecting microbial signatures is overcoming the high variability in microbial community composition both between and within hosts (or environmental habitats). Thus, transforming the raw data by collapsing or clustering the observed taxa or genes according to similarity is key. In the case of shotgun metagenomic sequences, we might first filter the sequences for known genes, and then assign them to functional or metabolic groups according to established databases prior to downstream analysis (Figure 15). For surveys of marker genes (such as 16S rRNA), we commonly cluster sequences into operational taxonomic units (OTUs) based on a pre-determined threshold of nucleotide similarity (e.g. 97%). However, when we perform data transformation as a fixed pre-processing step, we may be making incorrect assumptions about the best way to collapse input data for a given predictive task. Alternatively, we propose that the next generation of predictive models must be able to integrate external information sources into the process of feature selection to determine the appropriate levels of collapsing, filtering, or clustering.

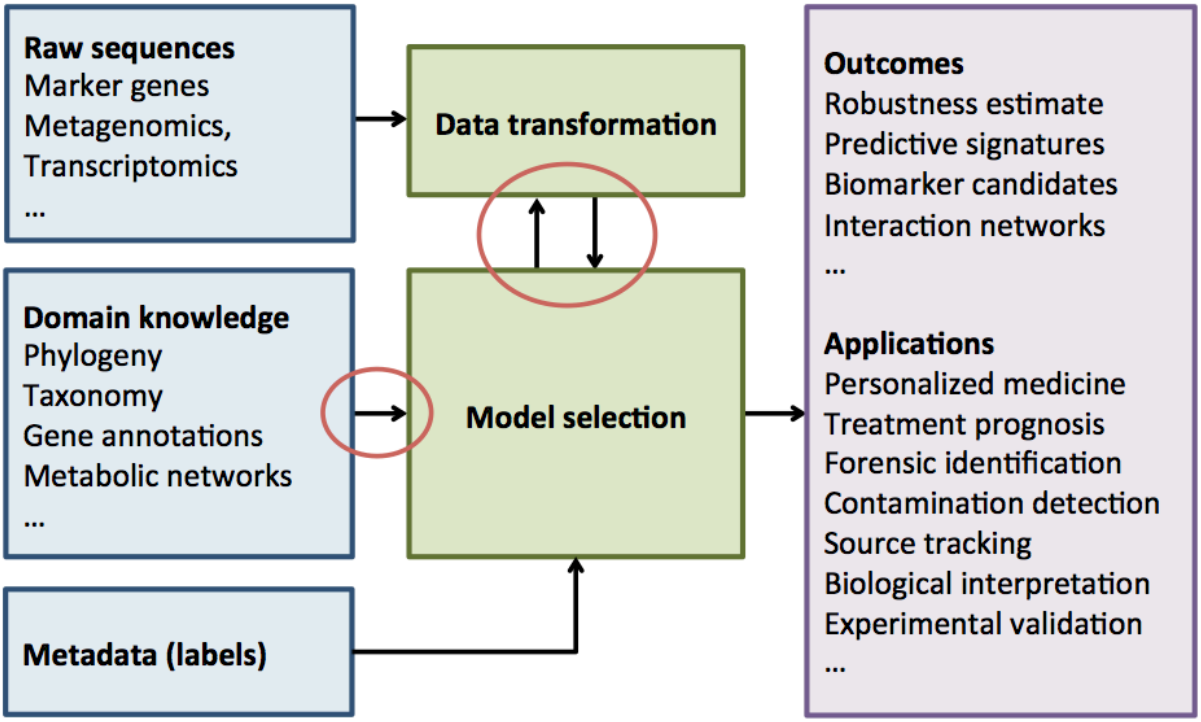

Figure 15. Processes for microbial signature discovery

The process begins with the collection of a large set of sequencing data from various bacterial communities associated with different environments or different host phenotypes. These sequences can serve directly as input to a machine learning algorithm, or they can be transformed through a preprocessing step (data transformation). Although for microbial community analysis data transformation and supervised learning are typically performed as separate steps, we suggest that predictive models will be improved by the development of novel machine learning techniques that are informed by the potential data transformations. For example, constructing a good predictive model using metabolic characterizations of metagenomics sequences might be easier if the algorithm has knowledge of the hierarchical relationships between metabolic functions. In the case of marker-gene surveys, a machine learning algorithm may benefit from knowledge of the phylogenetic relationships of the observed lineages, or the network of average nucleotide similarities between the input sequences. These structures may allow models to share statistical strength across related independent variables in cases where there is high variability within a given environment or host phenotype (i.e. lack of a "core microbiome").

For example, when we pick OTU clusters for marker-gene sequences at a fixed threshold, potentially discriminative taxa may lose their signal if we make the clusters either too specific (e.g. 99% similarity), or too broad (e.g. 80% similarity). In the case where the clusters are too specific, any conclusions made about those clusters may not generalize well to future data sets due to high variability between communities. This potential pitfall is referred to as “overfitting”. Many published studies use a within-cluster similarity threshold of 97%, but we have found that this is not usually the best level for predictive modeling. In the context of predictive modeling, it is possible to estimate the best OTU threshold empirically as the one that minimizes the expected future error of a classifier. We studied six human-associated microbial communities with well-understood clustering patterns to determine their optimal OTU thresholds for predictive modeling. Three examples are shown in Figure 16. For a given benchmark, we estimated the generalization error of the Random Forests classifier [6] using as input features OTUs picked at thresholds ranging from 60% to 99.5% nucleotide similarity. We then chose the optimal threshold for a given benchmark as the one giving the most parsimonious model (fewest OTUs) within one standard error of the best model. Optimal thresholds for the six tasks were surprisingly variable, ranging from 76% to 99%). This implies that predictive models are likely to benefit from a flexible approach to picking predictive OTU clusters, instead of the current practice of clustering at a fixed, pre-defined threshold of 97%.

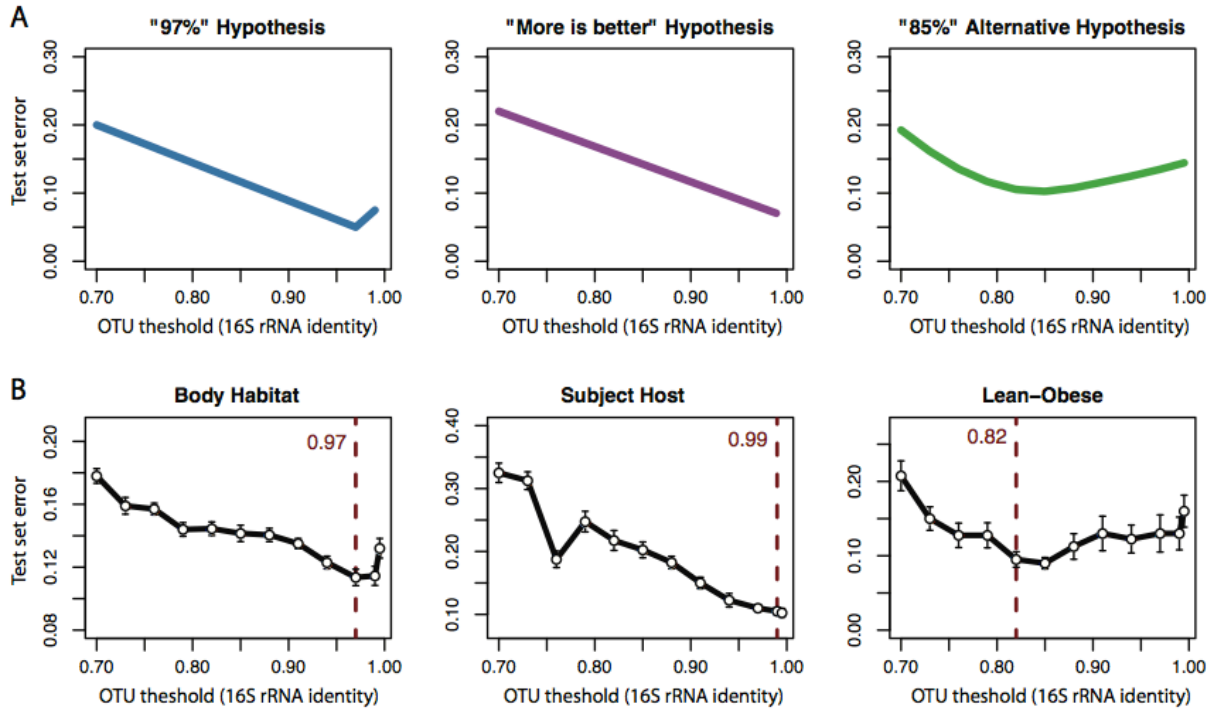

Figure 16. Are we overfitting with 97% OTUs?

Many microbial ecology studies use operational taxonomic units (OTUs) defined at 97% 16S SSU rRNA sequence identity, consistent with the conventional bacterial species threshold. However, it is possible that either more specific, or more general OTU definitions may be useful for machine learning studies. Panel A shows hypothetical error curves for the case that the commonly used 97% 16S SSU rRNA identity threshold represents an optimal OTU definition for a given classification task, the case that more specific OTUs are always better, and the case that the optimal identity threshold is lower, for example 85%. The hypothetical error curves illustrate the concepts of "overfitting" and "underfitting": if the clusters are too specific, then a predictive model cannot observe general trends in the data (overfitting); if they are too general, then the predictive features are getting buried during the clustering (underfitting). Panel B relates the choice of OTU threshold to empirical error in correctly classifying samples using a random forest classifier 4 trained on two-thirds of the data and tested on the remaining third, for 10 randomly chosen train/test splits of the data. Three classification benchmarks are shown: the Body Habitat benchmark categorizes host-associated microbial communities by general body habitat; the Host Subject benchmark categorizes communities from the forearm, palm, and index finger by host subject; the Lean-Obese benchmark categorizes gut communities by host phenotype. Vertical dashed lines indicate the most parsimonious model (i.e. fewest OTUs) whose mean generalization error is within one standard error of the best model. The empirical error curves (B) suggest that different classification tasks may be best accomplished with different OTU definitions. This is a demonstration of our more general suggestion that existing knowledge about raw input data, whether marker genes or shotgun metagenomic sequences, must be incorporated into the next generation of predictive algorithms.

Furthermore, a recent exploratory study found that several host quantitative trait loci influenced the relative abundance of taxonomic groups of variable breadth [83], indicating that even within a given classification task, a single threshold for taxonomic clustering may be insufficient to capture the relevant habitat-related adaptations of microbial communities.

For this reason, we believe that information about the nucleotide similarity or phylogenetic relationships of the input 16S rRNA sequences should be supplied directly to the machine learning algorithm, as shown in Figure 15. This will require the development of novel algorithms, but it has the benefit that the algorithm may select the appropriate levels of specificity for clustering input sequences given a particular predictive task. In the case of shotgun metagenomic sequences, we may cluster according to existing ontologies.

5.3 Biological considerations and validation

Assuming that we are able to identify microbial signatures that are predictive of, for example, a diseased host phenotype, it may still be difficult to determine whether differences in "discriminating" taxa are a cause or a consequence of disease without large prospective longitudinal studies. As an example, although the composition of the vaginal microbiota may impact the rate at which HIV is transmitted, subsequent changes to the vaginal microbiota due to immune-dysfunction would make it impossible to characterize a community signature that may pre-dispose an individual to HIV infection by comparing the vaginal microbiota of HIV positive women to healthy controls. Similarly, individuals with IBD and celiac disease are believed to have increased intestinal permeability prior to the onset of disease [84], and it is reasonable to expect that corresponding changes, such as alterations in the phospholipid composition in the intestinal mucous barrier [85], may be associated with characteristic changes in particular bacterial species (e.g. promoting

particular mucolytic species). Studies of how the microbiota differ with IBD, however, have generally compared people who have already developed the disease to those who have not [86]. Consequently, taxa that differ may be those that can tolerate inflammation in the gut, and not those that are causing it, or those whose presence could predict disease onset.

Assuming that microbial signatures can be successfully associated with host traits, there are still many issues of interpretation that complicate attempts to make biological or mechanistic conclusions from those associations. The most reliable microbial markers for hard-to-observe host conditions will be backed both by extensive correlation data across studies and well-understood mechanisms that relate phenotype to particular genes, organisms, or community features. Two particularly noteworthy approaches to supplementing correlation data with mechanism include experimental confirmation, and genomic studies of microbial lineages. As an example of the first approach, Sharon *et al.* [87] applied a combination of correlation studies and experimental confirmation to uncover a bacterium involved in *Drosophila melanogaster* mate preference. It had previously been observed that *Drosophila* raised on different media interbred less than those raised on the same medium. Investigation of the fly microbiota revealed that some lineages, in particular the *Lactobacilli*, differed in flies raised on different media, indicating that this could be either a cause or secondary marker of the observed difference in mate preferences. To distinguish between these possibilities, Sharon *et al.*, demonstrated that broad-spectrum antibiotics could abolish the observed mate preference. Adding *Lactobacillus plantarum* could rescue the mate preference effect in antibiotic-treated flies. Such experimental confirmation greatly strengthens the case for approaches that would seek to use *L. plantarum* levels as a marker for mate preference in wild Drosophila populations beyond what could be said from correlation data alone. Further characterization of the mechanism

involved in *L. plantarum* modification of mate preference (e.g. does it affect Drosophila pheromones?) would make this an even stronger candidate as a marker.

In cases where experimental manipulation is difficult, additional mechanistic information into the role of a putative marker microbe can be gained by examination of genome sequences. For example, Turnbaugh et al. [25] used a combination of genomic and transcriptomic approaches to study members of class Erysipelotrichi that increased when gnotobiotic mice, transplanted with a human microbial community, were switched from a low-fat diet rich in vegetables to a high-fat, high-sugar diet. These analyses found the genome of the cultured isolate to be enriched in phosphotransferase system (PTS) transporters, and identified PTS genes involved in the import of simple sugars as upregulated following the switch to a sucrose- and fat- rich western diet. Such genomic and transcriptomic findings supported the hypothesis that the observed increase in Erysipelotrichi was caused by changes in diet.

In some cases, models of human-associated microbial communities can already give reasonably accurate predictions of important traits such as host phenotype, forensic identification of the host [5], and environmental sources of sample contamination [3]. There is likely an enormous potential for improvement, however, with the increased availability of training data from a broad variety of prospective studies and the development of novel theoretical approaches that account for latent structures such as the phylogeny and behavioral characteristics of a microbiome. Experimental validation and biological interpretation of predictive models is also essential as the field moves toward high-stakes applications including personalized medicine and the early diagnosis of disease.

6 Bibliography

1. Knights D, Costello EK, Knight R: **Supervised classification of human microbiota**. *FEMS Microbiol Rev* 2011, **35**(2):343-359.
2. Knights D, Parfrey LW, Zaneveld J, Lozupone C, Knight R: **Human-associated microbial signatures: examining their predictive value**. *Cell Host & Microbe* 2011, **10**(4):292-296.
3. Knights D, Kuczynski J, Charlson ES, Zaneveld J, Mozer MC, Collman RG, Bushman FD, Knight R, Kelley ST: **Bayesian community-wide culture-independent microbial source tracking**. *Nat Methods* 2011, **8**(9):761-763.
4. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI: **The human microbiome project**. *Nature* 2007, **449**(7164):804-810.
5. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R: **Forensic identification using skin bacterial communities**. *Proc Natl Acad Sci U S A* 2010, **107**(14):6477-6481.
6. Breiman L: **Random Forests**. *Machine Learning* 2001, **45**(1):5-32.
7. Werner JJ, Knights D, Garcia ML, Scalfone NB, Smith S, Yarasheski K, Cummings TA, Beers AR, Knight R, Angenent LT: **Bacterial community structures are unique and resilient in full-scale bioenergy systems**. *Proc Natl Acad Sci U S A* 2011, **108**(10):4158-4163.
8. Muegge BD, Kuczynski J, Knights D, Clemente JC, González A, Fontana L, Henrissat B, Knight R, Gordon JI: **Diet Drives Convergence in Gut Microbiome Functions Across Mammalian Phylogeny and Within Humans**. *Science* 2011, **332**(6032):970-974.

9. McNulty NP, Yatsunenko T, Hsiao A, Faith JJ, Muegge BD, Goodman AL, Henrissat B, Oozeer R, Cools-Portier S, Gobert G *et al*: **The impact of a consortium of fermented milk strains on the gut microbiome of gnotobiotic mice and monozygotic twins**. *Science Translational Medicine* 2011, **3**(106):106ra106.

10. Smith A, Sterba-Boatwright B, Mott J: **Novel application of a statistical technique, Random Forests, in a bacterial source tracking study**. *Water Res* 2010, **44**(14):4067-4076.

11. Dufrêne M, Legendre P: **Species Assemblages and Indicator Species: The Need for a Flexible Asymmetrical Approach**. *Ecol Monogr* 1997, **67**(3):345-366.

12. Simpson JM, Santo Domingo JW, Reasoner DJ: **Microbial source tracking: state of the science**. *Environ Sci Technol* 2002, **36**(24):5279-5288.

13. Lozupone CA, Knight R: **Species divergence and the measurement of microbial diversity**. *FEMS Microbiol Rev* 2008, **32**(4):557-578.

14. Wen L, Ley RE, Volchkov PY, Stranges PB, Avanesyan L, Stonebraker AC, Hu C, Wong FS, Szot GL, Bluestone JA *et al*: **Innate immunity and intestinal microbiota in the development of Type 1 diabetes**. *Nature* 2008, **455**(7216):1109-1113.

15. Li M, Wang B, Zhang M, Rantalainen M, Wang S, Zhou H, Zhang Y, Shen J, Pang X, Zhang M *et al*: **Symbiotic gut microbes modulate human metabolic phenotypes**. *Proc Natl Acad Sci U S A* 2008, **105**(6):2117-2122.

16. Grice EA, Kong HH, Conlan S, Deming CB, Davis J, Young AC, Bouffard GG, Blakesley RW, Murray PR, Green ED *et al*: **Topographical and temporal diversity of the human skin microbiome**. *Science* 2009, **324**(5931):1190-1192.

17. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP *et al*: **A core gut microbiome in obese and lean twins**. *Nature* 2009, **457**(7228):480-484.

18. Magurran AE: **Measuring biological diversity**. Oxford: Blackwell Publishing; 2004.

19. Fierer N, Hamady M, Lauber CL, Knight R: **The influence of sex, handedness, and washing on the diversity of hand surface bacteria**. *Proc Natl Acad Sci U S A* 2008, **105**(46):17994-17999.

20. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R: **Bacterial community variation in human body habitats across space and time**. *Science* 2009, **326**(5960):1694-1697.

21. Martin AP: **Phylogenetic Approaches for Describing and Comparing the Diversity of Microbial Communities**. *Appl Environ Microbiol* 2002, **68**(8):3673-3682.

22. Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI: **Worlds within worlds: evolution of the vertebrate gut microbiota**. *Nat Rev Microbiol* 2008, **6**(10):776-788.

23. Wang Q, Garrity GM, Tiedje JM, Cole JR: **Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy**. *Appl Environ Microbiol* 2007, **73**(16):5261-5267.

24. Clayton TA, Baker D, Lindon JC, Everett JR, Nicholson JK: **Pharmacometabonomic identification of a significant host-microbiome metabolic interaction affecting human drug metabolism**. *Proc Natl Acad Sci U S A* 2009, **106**(34):14728-14733.

25. Turnbaugh PJ, Ridaura VK, Faith JJ, Rey FE, Knight R, Gordon JI: **The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice**. *Science translational medicine* 2009, **1**(6):6ra14-16ra14.

26. Hehemann J-H, Correc G, Barbeyron T, Helbert W, Czjzek M, Michel G: **Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota**. *Nature* 2010, **464**(7290):908-912.

27. Hooper LV: **Commensal Host-Bacterial Relationships in the Gut**. *Science* 2001, **292**(5519):1115-1118.

28. Schloss PD, Handelsman J: **Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness**. *Appl Environ Microbiol* 2005, **71**(3):1501-1506.

29. Lee JW, Lee JB, Park M, Song SH: **An extensive comparison of recent classification tools applied to microarray data**. *Computational Statistics & Data Analysis* 2005, **48**(4):869-885.

30. Hastie T, Tibshirani R, Friedman J: **The Elements of Statistical Learning, Second Edition: Data Mining, Inference, and Prediction**: Springer; 2009.

31. Blei DM, Ng AY, Jordan MI: **Latent dirichlet allocation**. *The Journal of Machine Learning* 2003.

32. Hinton GE, Osindero S, Teh Y-W: **A Fast Learning Algorithm for Deep Belief Nets**. *Neural Comp* 2006, **18**(7):1527-1554.

33. Yang C, Mills D, Mathee K, Wang Y, Jayachandran K, Sikaroodi M, Gillevet P, Entry J, Narasimhan G: **An ecoinformatics tool for microbial community studies: Supervised classification of Amplicon Length Heterogeneity (ALH) profiles of 16S rRNA**. *J Microbiol Methods* 2006, **65**(1):49-62.

34. Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ: **Random forests for classification in ecology**. *Ecology* 2007, **88**(11):2783-2792.

35. Quince C, Lanzén A, Curtis TP, Davenport RJ, Hall N, Head IM, Read LF, Sloan WT: **Accurate determination of microbial diversity from 454 pyrosequencing data**. *Nat Methods* 2009, **6**(9):639-641.

36. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI *et al*: **QIIME allows analysis of high-throughput community sequencing data**. *Nat Methods* 2010, **7**(5):335-336.

37. Edgar RC: **UCLUST**. In.: Robert C. Edgar; 2010.

38. Guyon I, Elisseeff A: **An Introduction to Variable and Feature Selection**. *Journal of Machine Learning Research* 2003, **3**:1157-1182.

39. Forman G: **An Extensive Empirical Study of Feature Selection Metrics for Text Classification**. *Journal of Machine Learning Research* 2003, **3**:1289-1305.

40. Man MZ, Dyson G, Johnson K, Liao B: **Evaluating methods for classifying expression data**. *J Biopharm Stat* 2004, **14**(4):1065-1084.

41. Saeys Y, Inza I, Larranaga P: **A review of feature selection techniques in bioinformatics**. *Bioinformatics* 2007, **23**(19):2507-2517.

42. Lal TN, Chapelle O, Weston J, Elisseeff A: **Embedded methods**. In. Edited by Guyon I, Gunn S, Nikravesh M, Zadeh LA. Berlin, Germany: Springer; 2006: 137-165.

43. Tibshirani R, Hastie T, Narasimhan B, Chu G: **Diagnosis of multiple cancer types by shrunken centroids of gene expression**. *PNAS* 2002, **99**(10):6567-6572.

44. Zou H, Hastie T: **Regularization and variable selection via the Elastic Net**. *Journal of the Royal Statistical Society B* 2005, **67**:301-320.

45. Gashler M, Giraud-Carrier C, Martinez T: **Decision tree ensemble: small heterogeneous is better than large homogeneous**. In*: 2008*. 900-905.

46. Cortes C, Vapnik V: **Support Vector Networks**. In*: 1995*. 273-297.

47. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S: **A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis**. *Bioinformatics* 2005, **21**(5):631-631.

48. Lee Y, Lin Y, Wahba G: **Multicategory Support Vector Machines: Theory and Application to the Classification of Microarray Data and Satellite Radiance Data**. *Journal of the American Statistical …* 2004.

49. Weston J, Mukherjee S, Chapelle O, Pontil M, Poggio T, Vapnik V: **Feature selection for SVMs**. *Advances in neural information processing systems* 2001:668-674.

50. Weston J, Elisseeff A, Scholkopf B, Tipping M, Kaelbling P: **Use of the Zero-Norm with Linear Models and Kernel Methods**. *Journal of Machine Learning Research* 2003, **3**:1439-1461.

51. Guyon I, Weston J, Barnhill S, Vapnik V: **Gene Selection for Cancer Classification using Support Vector Machines**. *Machine Learning* 2002, **46**(1):389-422.

52. Liaw A, Wiener M: **Classification and Regression by randomForest**. *R News* 2002, **2**(3):18-22.

53. Hastie T, Tibshirani R, Narasimhan B, Chu G: **pamr: prediction analysis for microarrays**. In.; 2009.

54. Friedman J, Hastie T, Tibshirani R: **Regularization Paths for Generalized Linear Models via Coordinate Descent**. *Journal of Statistical Software* 2010, **33**(1):1-22.

55. Kuhnert P, Christensen H: **Pasteurellaceae: Biology, Genomics and Molecular Aspects**: Horizon Scientific Press; 2008.

56. Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel A: **e1071: Misc Functions of the Department of Statistics (e1071), TU Wien**. In.; 2010.

57. Lozupone C, Knight R: **UniFrac: a New Phylogenetic Method for Comparing Microbial Communities**. *Appl Environ Microbiol* 2005, **71**(12):8228-8235.

58. Horner-Devine MC, Silver JM, Leibold MA, Bohannan BJM, Colwell RK, Fuhrman JA, Green JL, Kuske CR, Martiny JBH, Muyzer G *et al*: **A comparison of taxon co-occurrence patterns for macro- and microorganisms**. *Ecology* 2007, **88**(6):1345-1353.

59. McCallum A, Pal C, Wang X, Druck G: **Multi-conditional learning: Generative/discriminative training for clustering and classification**. 2006.

60. Blei DM, McAuliffe J: **Supervised topic models**. *Advances in Neural Information Processing …* 2008.

61. Chang J: **lda: Collapsed Gibbs sampling methods for topic models**. In.; 2010.

62. McCallum A, Nigam K: **A comparison of event models for naive bayes text classification**. In*: 1998*. Citeseer.

63. Lee SS: **Noisy replication in skewed binary classification**. *Computational statistics \& data analysis* 2000, **34**(2):165-191.

64. Nair V, Hinton G: **3D Object Recognition with Deep Belief Nets**. In. Edited by Bengio Y, Schuurmans D, Lafferty J, Williams CKI, Culotta A: MIT Press; 2009: 1339-1347.

65. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV *et al*: **The minimum information about a genome sequence (MIGS) specification**. *Nat Biotech* 2008, **26**(5):541-547.

66. De Filippo C, Cavalieri D, Di Paola M, Ramazzotti M, Poullet JB, Massart S, Collini S, Pieraccini G, Lionetti P: **Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa**. *Proc Natl Acad Sci U S A* 2010, **107**(33):14691-14696.

67. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM *et al*: **Enterotypes of the human gut microbiome**. *Nature* 2011, **473**(7346):174-180.

68. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R *et al*: **Linking long-term dietary patterns with gut microbial enterotypes**. *Science* 2011, **334**(6052):105-108.

69. Kaufman L, Rousseeuw PJ: **Finding groups in data : an introduction to cluster analysis**. New York: Wiley; 1990.

70. Rousseeuw PJ: **Silhouettes: a graphical aid to the interpretation and validation of cluster analysis**. *Journal of computational and applied mathematics* 1987, **20**:53-65.

71. Caliński T, Harabasz J: **A dendrite method for cluster analysis**. *Communications in Statistics-Theory and Methods* 1974, **3**(1):1-27.

72. Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF: **PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample**. *Appl Environ Microbiol* 2005, **71**(12):8966-8969.

73. Tanner MA, Goebel BM, Dojka MA, Pace NR: **Specific ribosomal DNA sequences from diverse environmental settings correlate with experimental contaminants**. *Appl Environ Microbiol* 1998, **64**(8):3110-3113.

74. Wu CH, Sercu B, Van de Werfhorst LC, Wong J, DeSantis TZ, Brodie EL, Hazen TC, Holden PA, Andersen GL: **Characterization of coastal urban watershed bacterial communities leads to alternative community-based indicators**. *PLoS One* 2010, **5**(6):e11285.

75. Greenberg J, Price B, Ware A: **Alternative estimate of source distribution in microbial source tracking using posterior probabilities**. *Water Res* 2010, **44**(8):2629-2637.

76. Lauber CL, Hamady M, Knight R, Fierer N: **Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale**. *Appl Environ Microbiol* 2009, **75**(15):5111-5120.

77. Griffiths TL, Steyvers M: **Finding scientific topics**. *Proc Natl Acad Sci U S A* 2004, **101**:5228-5235.

78. Wu GD, Lewis JD, Hoffmann C, Chen YY, Knight R, Bittinger K, Hwang J, Chen J, Berkowsky R, Nessel L *et al*: **Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags**. *BMC Microbiol* 2010, **10**:206.

79. Reeder J, Knight R: **Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions**. *Nat Methods* 2010, **7**(9):668-669.

80. Edgar RC: **Search and clustering orders of magnitude faster than BLAST**. *Bioinformatics* 2010, **26**(19):2460-2461.

81. Haas BJ, Gevers D, Earl A, Feldgarden M, Ward DV, Giannokous G, Ciulla D, Tabbaa D, Highlander SK, Sodergren E *et al*: **Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons**. *Genome Res* 2011.

82. Price MN, Dehal PS, Arkin AP: **FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix**. *Mol Biol Evol* 2009, **26**(7):1641-1650.

83. Benson AK, Kelly SA, Legge R, Ma F, Low SJ, Kim J, Zhang M, Oh PL, Nehrenberg D, Hua K *et al*: **Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors**. *Proc Natl Acad Sci U S A* 2010, **107**(44):18933-18938.

84. Groschwitz KR, Hogan SP: **Intestinal barrier function: molecular regulation and disease pathogenesis**. *The Journal of allergy and clinical immunology* 2009, **124**(1):3-20; quiz 21-22.

85. Braun A, Treede I, Gotthardt D, Tietje A, Zahn A, Ruhwald R, Schoenfeld U, Welsch T, Kienle P, Erben G *et al*: **Alterations of phospholipid concentration and species composition of the intestinal mucus barrier in ulcerative colitis: a clue to pathogenesis**. *Inflamm Bowel Dis* 2009, **15**(11):1705-1720.

86. Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR: **Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases**. *Proc Natl Acad Sci U S A* 2007, **104**(34):13780-13785.

87. Sharon G, Segal D, Ringo JM, Hefetz A, Zilber-Rosenberg I, Rosenberg E: **Commensal bacteria play a role in mating preference of Drosophila melanogaster**. *Proc Natl Acad Sci U S A* 2010, **107**(46):20051-20056.